N° d'ordre : 37

# THÈSE

PRÉSENTÉE

A LA FACULTÉ DES SCIENCES DE L'UNIVERSITÉ DE BESANÇON

POUR OBTENIR

LE TITRE DE DOCTEUR EN MATHÉMATIQUES

Mention : MATHÉMATIQUES PURES

PAR

**Dominique GROSSIORD**

Licencié ès-Sciences

---

# THÈSE

## APPLICATION DE LA MÉTHODE DE LAGUERRE AU CALCUL DES VALEURS PROPRES DE MATRICES TRIDIAGONALES

---

Soutenue le 1er Juin 1966, devant la Commission d'Examen

JURY : MM. **CHATELET,** *Professeur* — Président
**RIGAL,** *Professeur* — Examinateurs
**PARIZET,** Maître de Conférence — Examinateurs

N° d'ordre : 37

# THÈSE

PRÉSENTÉE

A LA FACULTÉ DES SCIENCES DE L'UNIVERSITÉ DE BESANÇON

POUR OBTENIR

LE TITRE DE DOCTEUR EN MATHÉMATIQUES

Mention : MATHÉMATIQUES PURES

PAR

**Dominique GROSSIORD**

Licencié ès-Sciences

---

# THÈSE

## APPLICATION DE LA MÉTHODE DE LAGUERRE AU CALCUL DES VALEURS PROPRES DE MATRICES TRIDIAGONALES

---

Soutenue le 1er Juin 1966, devant la Commission d'Examen

JURY : MM. **CHATELET,** *Professeur* — Président
**RIGAL,** *Professeur* — Examinateurs
**PARIZET,** Maître de Conférence — Examinateurs

UNIV.
LEIDEN
BIBL.

A Monsieur Robert CROISOT

Professeur de Mathématiques à la Faculté des Sciences

décédé accidentellement le 1 Avril 1966

Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41547626_0013

Je tiens à assurer de toute ma reconnaissance :

Monsieur le Professeur CHATELET dont j'ai pu bénéficier de l'enseignement et qui a bien voulu accepter la présidence de mon jury.

Monsieur le Professeur RIGAL, Directeur du Laboratoire d'Informatique de la Faculté des Sciences qui, dès la création de son Laboratoire, m'a initié à l'Analyse Numérique et qui, en acceptant la Direction de cette thèse, n'a cessé de me prodiguer conseils et encouragements durant les années de préparation.

Monsieur PARIZET, Maître de Conférence, qui m'a fait l'honneur de bien vouloir faire partie de mon jury.

Je remercie Monsieur le Professeur CASTINEL de l'Institut de Mathématiques Appliquées de Grenoble de m'avoir offert, avec ses conseils, la possibilité d'utiliser l'I.B.M. 7044 de son Laboratoire.

Je remercie aussi tous les membres du Laboratoire qui m'ont aidé, notamment lors des essais sur CAB 500 et les secrétaires pour le soin apporté à la présentation matérielle de ce travail.

---

SOMMAIRE

Le CHAPITRE I rappelle ce que sont les conventions arithmétiques et précise la notion d'indiscernabilité de deux nombres-machines. Parmi les divers estimateurs d'écart qui sont présentés, nous introduisons en particulier les notions de distance relative et d'écart relatif qui permettent de préciser la signification des opérations élémentaires ( $\oplus$ , $\ominus$ , $\otimes$ , $\emptyset$, produit scalaire) en "virgule flottante normalisée". Le programme de troncature est une première approche de l'estimation d'une marge d'erreur mais nous indiquons, en conclusion, la possibilité de réaliser d'autres arithmétiques estimant mieux les erreurs de calcul.

Le CHAPITRE II présente la méthode de Laguerre (déduite de celle de Newton) sous deux formes : la méthode $\mathcal{L}_\infty$, la méthode $\mathcal{L}_n$. Nous précisons dans chaque cas et comparons leurs propriétés, notamment la nature de la convergence (généralement monotone et cubique) vers une solution réelle simple ou multiple. L'application de cette méthode au calcul des solutions réelles d'un polynôme conduit à définir un algorithme original utilisant au mieux les propriétés propres à la méthode de Laguerre : Un mode spécial de balayage s'est révélé particulièrement puissant et efficace, en liaison avec des tests d'arrêts convenables définis à partir de l'étude de la zone d'incertitude (zone où aucune information significative sur les solutions ne peut être obtenue).

Le CHAPITRE III étudie les applications de la méthode de Laguerre au calcul des valeurs propres d'une matrice tridiagonale. Après avoir indiqué de quelles manières pouvaient être calculées le polynôme caractéristique, ses dérivées et ses coefficients, nous présentons quelques propriétés propres à ce type de matrice et introduisons les notions de matrices symétrisables et réduites. Le double point de vue, polynomial ou matriciel que l'on a pu développer en calculant le polynôme caractéristique ou en localisant les valeurs propres conduit à définir deux algorithmes qui seront critiqués.

Le CHAPITRE IV regroupe plusieurs ensembles de matrices-test tridiagonales à valeurs propres réelles, proposées par divers auteurs, et étudie leurs valeurs propres. Nous donnons en annexe le calcul analytique des valeurs propres de certaines matrices et énonçons un théorème sur les valeurs propres de matrices tridiagonales à termes diagonaux constants.

Le CHAPITRE V présente un ensemble de procédures ALGOL que nous avons utilisé et une partie des essais effectués sur CAB 500 ou sur I.B.M. 7044 qui permettent de signaler d'éventuelles améliorations de l'algorithme existant ou de montrer combien IL EST ILLUSOIRE DE VOULOIR CALCULER DES VALEURS PROPRES, LORSQUE, DANS LEUR VOISINAGE, LE POLYNOME CARACTERISTIQUE NE PEUT ETRE DETERMINÉ AVEC PRECISION.

# TABLE DES MATIERES

CHAPITRE I -

ARITHMETIQUES ET COMPARAISONS
PORTANT SUR DES NOMBRES-MACHINE

1.1 Les Conventions arithmétiques, application du Réel $\mathbb{R}$ sur un réel machine $\mathbb{R}^*$ ..... I - 2

1.2 La notion d'indiscernabilité, relation d'équivalence dans $\mathbb{R}^*$..... I - 5

1.3 La distance relative entre réels... I - 6

1.4 Quelques autres estimateurs de comparaison entre nombres réels.... I - 8

1.5 La différence relative entre nombres-machines..... I - 9

1.6 L'écart relatif entre nombres-machine et entre réels..... I - 10

1.7 Signification de l'addition et de la soustraction de deux nombres-machine..... I - 16

1.8 Remarques sur la multiplication, le produit scalaire et la division. I - 19

1.9 Programme de troncature ou simulateur de précision variable..... I - 22

Conclusion du Chapitre I..... I - 26

Réalisation du Programme de Troncature (Annexe)..... I - 28

Programme de Calcul de l'écart relatif (Annexe)..... I - 36

CHAPITRE II -

LA METHODE DE LAGUERRE


2.1 La méthode de Newton..... II - 1

2.2 La méthode $\mathcal{L}_\infty$ de Laguerre..... II - 2

2.3 La méthode $\mathcal{L}_n$ de Laguerre..... II - 6

2.4 Cas des solutions réelles multiples..... II - 13

2.5 Calcul numérique d'une solution d'un polynôme à solutions réelles selon la méthode de Laguerre..... II - 16

2.6 Algorithmes pratiques de calcul des solutions d'un polynôme..... II - 23


CHAPITRE III -

APPLICATION DE LA METHODE DE LAGUERRE AU CALCUL DES VALEURS PROPRES D'UNE MATRICE TRIDIAGONALE


3.1 Calcul du polynôme caractéristique et de ses dérivées successives.... III - 3

3.2 Propriétés des matières symétrisables..... III - 7

3.3 Localisation des valeurs propres.. III - 12

3.4 Calcul des vecteurs propres d'une matrice trigiadonale..... III - 15

3.5 Comparaison des points de vue polynomial et matriciel..... III - 22

CHAPITRE IV -

LISTES DE MATRICES - TEST

LISTE ..... IV - 1

Calcul analytique des valeurs propres d'une matrice de Wolstenholme et de quelques matrices déduites (Annexe).... IV - 11

CHAPITRE V -

PROGRAMME ET ESSAIS REALISES

5.1 Programme ALGOL..... V - 1

5.2 Quelques essais réalisés..... V - 17

5.3 Conclusion du Chapitre V..... V - 26

# CHAPITRE I

## Arithmétiques et Comparaisons portant sur des NOMBRES-MACHINE

*Réaliser un calcul numérique, c'est effectuer des opérations spéciales dites "OPERATIONS-MACHINE" et précisées par l'ensemble des microprogrammes, dans un ordre que définit l'ALGORITHME imposé par le mathématicien, sur les éléments appelés "NOMBRES MACHINE" d'un ensemble dit "REEL-MACHINE".*

*L'algorithme est finalement un opérateur, qui, portant sur une suite finie de nombres-machine (dite "DONNEES-MACHINE"), aboutit à une autre suite finie (dite "RESULTATS-MACHINE").*

*Cette suite peut comporter des réponses telles que "impossible", "indéterminé"...*

Il est théoriquement facile de définir, dans chaque algorithme, l'ensemble des opérations-machine, , celui des nombres-machine celui des données-machine et celui des résultats-machine.

Un même calcul peut faire appel-successivement ou simultanément- à diverses arithmétiques et à diverses représentations, le

passage des unes aux autres s'effectuant sous la tutelle de l'algorithme. Ce n'est pas ici le lieu de préciser plus mais d'insister sur les conséquences du fait que l'on travaille avec un ordinateur : CELUI-CI UTILISE NECESSAIREMENT DES NOMBRES, QUI NE SONT PAS DES NOMBRES REELS, mais des nombres-machine susceptibles d'un nombre fini d'états distincts.

## LES CONVENTIONS ARITHMETIQUES, APPLICATIONS DU REEL R SUR UN REEL-MACHINE $R^*$.

*On appellera convention-machine, toute application du réel R sur un ensemble $R^*$, dit réel-machine, dont les éléments sont précisément les nombres-machine, images des nombres réels.*

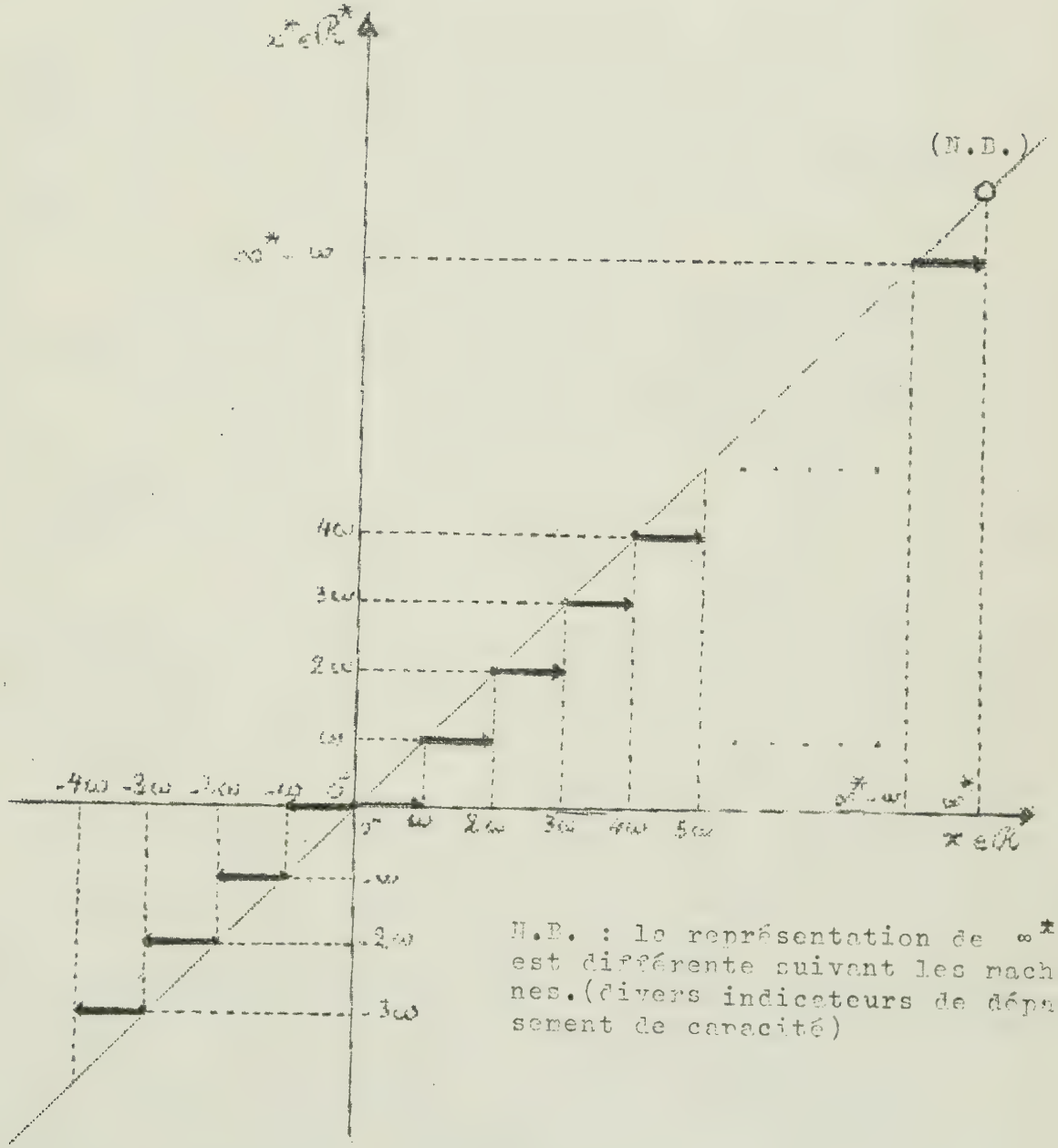
L'ENSEMBLE $R^*$ EST TOUJOURS FINI. Le nombre de ses éléments dépend de l'arithmétique utilisée : il est généralement de la forme $B^k$ où k est le nombre de "chiffres" utilisés dans le traitement par cette arithmétique (à l'exclusion de ceux utilisés par divers circuits technologiques tels que ceux du contrôle) et B (dit "base") est le nombre d'états distincts susceptibles d'être pris par chacun de ces chiffres. Pour des raisons technologiques et pour optimiser décemment la représentation et le traitement des nombres-machine, on choisit usuellement la base $B = 2$ (système binaire). Certains ordinateurs ont cependant été conçus en base $B = 3$ qui permet un emmagasinage optimal de l'information, d'autres en base $B = 10$ (système décimal dont la lecture nous est familière) ou en base $B = 8$ (système octal, variante du binaire plus commode à manier) ; d'autres encore utilisent plusieurs bases mais la conversion d'un nombre d'une base dans une autre, notamment lors des procédures de lecture et d'impression, modifie assez profondément l'information et la détériore partiellement.

L'ordinateur CAB 500, que nous avons utilisé, a un fonctionnement interne binaire et il est muni normalement de lecture et d'impression décimales. Nous avons écrit des microprogrammes spéciaux pour améliorer les procédures d'entrées et de sorties, par exemple des microprogrammes d'impression octale permettant de suivre l'évolution de calculs soumis à troncature ou de vérifier la notion "d'écart relatif" entre deux nombres-machine, points sur lesquels nous reviendrons.

FIGURE (1)

Convention virgule flottante sans arrondi

Fig(1)




N.B. : la représentation de $\omega^*$ est différente suivant les machines.(divers indicateurs de dépassement de capacité)

Légende de la Figure 1 -

*LA CONVENTION DITE "VIRGULE FIXE" SANS ARRONDI*

$\forall\ a \neq 0 \quad \left[|a| < \infty^*\right], \quad \exists\ a^* = K\omega$ (K entier algébrique : $|K| < 2^k$) : $|a^*| \leq |a| < |a^* + \omega|$

k est une constante (longueur des "mots") ainsi que ω dépendant de la "position de la virgule" telle que la précise le programme par l'intermédiaire du format (cf. infra note 2). ω (plus petit nombre discernable de zéro) est dit "minimal" ; $\infty^*$ (plus petit réel non représentable par suite de dépassement de capacité) est dit "maximal".

NOTONS LA RESSEMBLANCE DE CE SCHEMA AVEC CELUI RELATIF A TOUT RESULTAT EXPERIMENTAL : ω EST LA SENSIBILITE, $\infty^*$ CORRESPOND A LA SATURATION. ICI LE SEUIL DIFFERENTIEL EST CONSTANT.

x uniformément réparti d'une manière aléatoire entre 0 et $\infty^*$

$$\Longrightarrow \overleftarrow{x - x^*} = \frac{\omega}{2} \ ; \quad \sigma(x - x^*) = \frac{\omega}{\sqrt{3}}$$

*LA CONVENTION DITE "VIRGULE FIXE" AVEC ARRONDI* donne lieu à des remarques analogues. Mais le plus petit nombre discernable de zéro est $\frac{\omega}{2}$ . $\overleftarrow{x - x^*} = 0 \qquad \sigma(x - x^*) = \frac{\omega}{\sqrt{12}}$

Note 1 - En fait, la représentation est encore compliquée par les conventions relatives au signe (complémentarisation). Peu importe ici les fins détails technologiques ni le fait que $0 \in R$ ait 2 représentations ($0^+$ et $0^-$).

Note 2 - Le Format indique le nombre de chiffres significatifs avant et après la virgule. Il fixe la précision utilisée et permet de déterminer le maximal et le minimal.

Note 3 - Il sera souvent fait usage de cette convention pour représenter les nombres fractionnaires, en module inférieurs à 1. Alors $\infty^* = 1$.

Par une mise en facteur appropriée, il est en général possible de ramener tout calcul au traitement de tels nombres. Mais si l'on ne connaît pas d'une manière précise l'ordre de grandeur des données-machine, on sera tenté de travailler sur des nombres petits (devant 1) pour éviter les dépassements de capacité, limitant ainsi arbitrairement leur nombre de chiffres significatifs et, en conséquence, leur précision par suite des "zéros de garde" que l'on crée systématiquement à gauche.

## I.2. - LA NOTION D'INDISCERNABILITE, RELATION D'EQUIVALENCE DANS $R^*$

Définition

Deux réels sont indiscernables s'ils ont la même représentation machine. La relation d'équivalence ainsi définie n'est pas une relation de voisinage et ne se traduit pas en terme de distance. Deux réels aussi voisins que l'on veut (par exemple, en binaire, 0,01111... et 0,1000...) peuvent être discernables, ce qui est une conséquence du caractère discontinu de toute représentation arithmétique. Par contre, il est exact que deux nombres indiscernables (du moins en virgule fixe et si aucun des deux n'est représentable par $\infty^*$) sont distants de moins de $\omega$ l'un de l'autre (figure 1).

Remarque : la démonstration du fait que la relation d'indiscernabilité est une relation d'équivalence suppose pour tout réel l'existence d'un représentant unique ; cela peut être inexact pour certaines machines (comme celle donnant lieu à la représentation décrite par la figure 1) ayant deux "zéros-machine", l'un positif et l'autre négatif, susceptibles tous deux de représenter le même réel zéro.

Plus généralement, on est amené (par exemple RIGAL (1)) à définir des lieux d'incertitude sur les données-machine considérées comme suite de N réels-machine : ces lieux sont convexes mais non nécessairement hypercubiques.

## 1.3. - LA DISTANCE RELATIVE ENTRE REELS

On appellera "distance relative de deux réels", la fonction :

$$d(A,B) = \frac{|A - B|}{\max(|A|,|B|)} \quad ; \qquad d(0,0) = 0$$

Théorème 1.3.1 : la distance relative est une distance. Elle est invariante par homothétie mais non par translation.

Axiome 1 : $d(A,B) \geq 0$

Axiome 2 : $d(A,B) = 0 \Longleftrightarrow A = B$

Axiome 3 : $d(A,B) = d(B,A)$

Ces axiomes sont évidents, que A et B soient nuls ou non. Il en est de même des propriétés suivantes :

Propriété 5 : $d(CA, CB) = d(A,B) \quad \Longleftarrow (C \neq 0)$

Propriété 6 : $d(A + D, B + D) \neq d(A,B) \Longleftarrow (D \neq 0)$

Reste donc à démontrer l'axiome 4 (inégalité triangulaire)

$d(A,B) \leq d(A,C) + d(B,C)$

$$\frac{|A-B|}{\max(|A|,|B|)} \leq \frac{|A-C|}{\max(|A|,|C|)} + \frac{|B-C|}{\max(|B|,|C|)}$$

ce que nous allons faire, en supposant que ni A, ni B, ni C ne sont nuls (sinon la démonstration est immédiate). On distinguera plusieurs cas de figures ce qui est une démonstration lourde mais efficace. Nous pouvons commencer par supposer $|A| \leq |B|$ (ce qui est toujours possible compte tenu de la symétrie des rôles de A et B), et B positif, compte tenu de la propriété 5 et de la non-nullité postulée pour B. On est ramené à 3 cas différents :

. Cas 1 : $0 < |C| \leq |A| \leq |B|$

$$|A-B| \leq |A-C| + |C-B| \Longrightarrow \frac{|A-B|}{|B|} \leq \frac{|A-C|}{|B|} + \frac{|B-C|}{|B|} \leq \frac{|A-C|}{|A|} + \frac{|B-C|}{|B|}$$

$$\Longleftrightarrow \quad d(A,B) \leq d(A,C) + d(B,C).$$

. Cas 2 : $0 < |A| \leq |C| \leq |B|$

On utilise une démonstration analogue

$$d(A,B) = \frac{|A-B|}{|B|} \leq \frac{|A-C|}{|B|} + \frac{|B-C|}{|B|} \leq \frac{|A-C|}{|C|} + \frac{|B-C|}{|B|} = d(A,C) + d(B,C)$$

ou encore (B est toujours supposé positif)

$$\frac{|B-A|}{|B|} \leq \frac{|C-A|}{|C|} + \frac{|B-C|}{|B|} \Longleftrightarrow |C|\,(B-A) \leq B|C-A| + |C|(B-C)$$

$$\Longleftrightarrow |C|\,(C-A) \leq B|C-A|$$

cette dernière relation est vraie quelque soit le signe de C

. Cas 3 : $0 < |A| \leq |B| \leq |C|$

$$\frac{|B-A|}{|B|} \leq \frac{|C-A|}{|C|} + \frac{|C-B|}{|C|} \Longleftrightarrow \frac{|B-A|}{|B|} \leq \frac{|C-A|}{|C|} + \frac{|C-B|}{|C|} \Longleftrightarrow$$

$$\begin{cases} \alpha & [C>0 \quad C-B>0]\ C(B-A) \leq B(2C-A-B) \leftrightarrow -A(C-B) \leq B(C-B) \leftrightarrow -A \leq B \\ \beta & [C<0 \quad B-C>0]\ C(B-A) \geq B(2C-A-B) \leftrightarrow -A(B-C) \leq B(B-C) \leftrightarrow -A \leq B \end{cases}$$

Notons aussi les propriétés suivantes :

Propriété 7 : $0 \leq d(A,B) \leq 2$

Propriété 8 : $d(A,B) < 1 \Longleftrightarrow AB \geq 0$

$d(A,B) > 1 \Longleftrightarrow AB < 0$

Propriété 9 : $d(A,B) = 1 \Longleftrightarrow A=0,\ B\neq 0$ ou $A\neq 0,\ B=0$

: $d(A,B) = 2 \Longleftrightarrow A = -B \neq 0$

**Propriété 10:** $d(|A|,|B|) = d(A,B) \Longleftrightarrow AB \geq 0$

$d(|A|,|B|) = 2 - d(A,B) \Longleftrightarrow AB < 0$

## 1.4. QUELQUES AUTRES ESTIMATEURS DE COMPARAISON ENTRE NOMBRES REELS

Bien entendu il existe d'autres estimateurs, et en particulier les deux suivants :

. $d_2(A,B) = |A-B|\ B^{-\max(\alpha,\beta)}$ (seconde distance relative)

$d_2(0,0) = 0$

que nous définirons et étudierons bientôt (On remplace au dénominateur $|A|$ et $|B|$ par leur "ordre de grandeur" $B^\alpha$ et $B^\beta$, $\alpha$ et $\beta$ étant les exposants de A et B). Elle est équivalente (au sens des normes) à la distance relative (puisque $B^{\alpha-1} \leq A < B^\alpha$ et de même pour B).

. $d_3(A,B) = \dfrac{|A-B|}{|A| + |B|}$ (troisième distance relative)

$d_3(0,0) = 0$

dont nous nous bornons ici à préciser certaines propriétés

$d_3(A,B) = 1 \iff AB \leq 0$ (sauf A=B=0)

$d_3(A,B) < 1 \iff AB > 0$

$d_3(A,B) = 0 \iff A = B$

$d_3(CA,CB) = d_3(A,B) \quad C \neq 0$

$d_3$ est une distance : la démonstration de l'inégalité triangulaire est immédiate, sauf dans le cas où A, B, C sont de même signe, par exemple strictement positifs tous les trois. Dans ce cas, il faut toujours considérer 3 cas qui s'écrivent pour $0<A<B$ :

Cas 1 : $0 < C \leq A \leq B$

$$d_3(A,B) = \frac{B-A}{B+A} \leq \frac{A-C}{B+A} + \frac{B-C}{B+A} \leq \frac{A-C}{A+C} + \frac{B-C}{B+C} = d_3(A,C) + d_3(B,C)$$

Cas 2 : $0 < A \leq C \leq B$

$$d_3(A,B) = \frac{B-A}{B+A} \leq \frac{C-A}{A+C} + \frac{B-C}{B+C} \qquad (A+C)(B+C)(C-A) \leq$$

$$(B+A)\left[(C-A)(B+C)+(B-C)(A+C)\right] \iff AB+C^2 \leq BC+CA \iff C^2-C(A+B)+AB \leq 0$$

puisque C est précisément entre les racines (A et B) du trinôme.

Cas 3 : $0 < A \leq B \leq C$

$$\frac{B-A}{B+A} \leq \frac{C-A}{C+A} + \frac{C-B}{C+B} \qquad (C^2+CA+BC+AB) \leq 2(B+A)(C^2-AB)$$

$$3(C^2-B^2)A + (C^2-B^2)B + (C-B)A^2 \geq 0 \qquad C \geq B$$

. Pour A et B différents de zéro il existe les relations suivantes entre ces 3 distances

$$d_2(A,B) < d_1(A,B) \leq \mathcal{B}\ d_2(A,B)$$

$$\mathcal{B}^{\max(\alpha,\beta)-1} \leq \max(|A|,|B|) < \max(\alpha,\beta)$$

$$\frac{d_1(A,B)}{2} \leq d_3(A,B) \leq d_1(A,B)$$

$$\max(|A|,|B|) \leq |A|+|B| \leq 2 \max(|A|,|B|)$$

d'où

$$\frac{d_2(A,B)}{2} < \frac{d_1(A,B)}{2} \leq d_3(A,B) \leq d_1(A,B) \leq \mathcal{B}\ d_2(A,B)$$

## 1.5. LA DIFFERENCE RELATIVE ENTRE NOMBRES-MACHINE

Lorsque A et B sont des nombres-machine $A^*$ et $B^*$, la distance :

$$d(A^*, B^*) = \frac{|A^* - B^*|}{\max(|A^*|,|B^*|)}$$

définie par restriction sur le sous ensemble $R^*$, jouit de toutes les propriétés de la distance relative, mais elle n'est pas un nombre-machine. On peut alors convenir de définir et d'étudier la représentation machine de cette distance, où l'on aura remplacé les opérations arithmétiques par les opérations-machine correspondantes.

(Nous noterons celles-ci en entourant les opérateurs d'un rond ou d'un carré suivant qu'il s'agira des conventions "virgule fixe" ou "virgule flottante"), par exemple :

$$d^{\boxed{*}}(A^*, B^*) = |A^* \boxed{-} B^*| \quad \boxed{:} \max(|A^*|, |B^*|)$$

Par suite des erreurs de calcul, l'estimateur $d^{\boxed{*}}$ (que nous désignerons par la suite par $d^*$ car aucune confusion ne sera possible avec $d^{\circledast}$ que nous n'étudierons pas) n'est pas nécessairement le nombre-machine correspondant à $d(A^*, B^*)$.

*Il n'est pas du type"distance"* : il peut, par exemple, être nul lorsque $A^*$ et $B^*$ sont discernables :
$d^*(0{,}10000\ 10^{-98}, 0{,}9999\ 10^{-99})=0$ $A-B=0{,}00001\ 10^{-98} = 0$ si nous supposons que -99 est l'exposant minimal de la convention virgule flottante décimale utilisée dans cet exemple.

## 1.6. L'ECART RELATIF ENTRE NOMBRES-MACHINE ET ENTRE REELS.

*Cette difficulté théorique et le coût excessif d'obtention de la différence relative -et à plus forte raison des autres estimateurs- correspondant aux seconde et troisième distances relatives nous a conduit à une autre fonction, particulièrement adaptée à la convention dite "virgule flottante normalisée". L'ECART RELATIF, que nous allons définir, est équivalent à la distance relative. Il est en relation avec la notion intuitive de "nombres de chiffres significatifs communs à deux nombres écrits selon une même convention flottante normalisée" et celles, plus élaborée et sur laquelle nous reviendrons dans ce chapitre lors de l'étude des opérateurs-machine et des opérations correspondantes, de perte d'information et de création de "bruit" par suite de tout traitement-machine.*

Un nombre réel (A) est représenté en virgule flottante par un exposant que nous désignerons toujours par la suite, par la lettre grecque ($\alpha$) et par une mantisse ($a$)

$$A \rightarrow \{a,\alpha\}$$

La normalisation se traduit par la relation :

$$B^{-1} \leq |a| < 1 \qquad a \neq 0$$

qui précise le couple $(a,\alpha)$ pour tout A non nul en liaison avec la relation

$$A \rightarrow A^{*} = a\, B^{\alpha}.$$

Etant donné deux nombres réels, par exemple $0{,}12345.10^{6}$ et $0{,}12356.10^{6}$, il est courant de dire qu'ils ont 3 chiffres significatifs communs, ce qui signifie qu'ils sont "proches à $10^{-4}$ près". Mais cette notion intuitive est délicate à manier par suite du caractère discontinu de toute écriture arithmétique (0,9999 et 1,0000, bien que voisins, n'ont aucun chiffre significatif commun). C'est pour éliminer ce phénomène que nous partirons de la différence arithmétique pour définir un nombre-machine susceptible de représenter la différence relative entre deux nombres-machine.

Utilisant la "seconde différence relative" entre nombres-machine :

$$d_2(A^{*}, B^{*}) = |A^{*} - B^{*}| \quad B^{-\max(\alpha,\beta)}$$

$$d_2(0,0) = 0 \quad ; \quad d_2(A,0) = d_2(0,A) = |A^{*}| \quad B^{-\alpha} = |a|$$

Nous appelerons "écart relatif de deux réels", le représentant $d_2^*$ de la seconde différence relative de leurs représentants machine

$$e(A,B) = d_2^* \ (A^*,B^*)$$

que l'on calcule aisément à partir de l'une ou l'autre des formules suivantes :

$$(|A| \geq |B| \Longleftrightarrow \alpha \geq \beta) \Longrightarrow e(A,B) = |a \boxminus b \ \ B^{\beta-\alpha}|$$

$$(|A| \leq |B| \Longleftrightarrow \alpha \leq \beta) \Longrightarrow e(A,B) = |b \boxminus a \ \ B^{\beta-\alpha}|$$

Cette dernière expression montre que, lors du traitement effectué par le microprogramme addition-soustraction, l'ordinateur est nécessairement amené à former e(A,B) ou e(A,-B). Le programme correspondant, que nous avons écrit pour des nombres binaires (B=2) en CAB 500, est disponible et sera explicité en annexe.

Note 1 : $\alpha = \beta \Longrightarrow e(A,B) = |a - b|$

Note 2 : Bien entendu, les multiplications par une puissance négative de B s'opèrent par décalage à droite. Ceci nous oblige à conserver en toute rigueur le symbole $\boxminus$ à droite des formules précédentes (sauf si $\alpha=\beta$ ) et se traduit par une "perte d'information", conséquence de la disparition des $|\beta-\alpha|$ unités d'information situées à l'extrême droite de la représentation de $\min(|A|,|B|)$.

Théorème I : *L'écart relatif e(A,B) est une fonction symétrique et non négative de A et B. Il est nul si et seulement si les nombres A et B sont indiscernables.*

$$A \text{ et } B \text{ indiscernables} \Longleftrightarrow e(A,B)=0 \Longleftrightarrow A^* = B^*$$

Ce théorème fondamental résulte directement de la convention de normalisation adoptée pour la représentation flottante utilisée.

Théorème II : *L'écart relatif n'est pas du type distance.*

Il ne satisfait pas l'inégalité triangulaire

$e(A,B) \leqslant e(A,C) + e(B,C)$

que sous la réserve : $|C| < \max(|A|,|B|)$.

Dans ce cas, on peut écrire :

$$e(A,B) = |a-bB^{\beta-\alpha}| = |a-c\,B^{\gamma-\alpha} - B^{\beta-\alpha}\,(b-c\,B^{\gamma-\beta})|$$

$$\leqslant |a-c\,B^{\gamma-\alpha}| + B^{\beta-\alpha}\,|b-c\,B^{\gamma-\beta}|$$

$$\leqslant |a-c\,B^{\gamma-\alpha}| + |b-c\,B^{\gamma-\beta}| = e(A,C) + e(B,C)$$

Nous avons supposé ici $\gamma \leqslant \beta \leqslant \alpha$ mais une démonstration analogue est réalisable pour les autres possibilités.

Par contre, on démontre que l'inégalité n'est pas vérifiée lorsque $|C|$ est suffisamment grand en remarquant :

$$C \longrightarrow \infty \Longrightarrow \begin{cases} e(A,C) \rightarrow |c| \\ e(B,C) \rightarrow |c| \end{cases}$$

il est toujours possible de choisir A, B et C tels que le premier membre e(A,B) soit voisin de 2 et le second (e(A,C)+e(B,C) ) de $2\,B^{-1}$, quantité inférieure à la précédente.

L'opérateur écart relatif n'est invariant ni par homothétie, ni par translation. Nous pouvons seulement affirmer que (le cas de dépassement de capacité, par suite d'une valeur excessive de k, étant exclu) :

$$e(B^k A^*, B^k B^*) = e(A^*, B^*) = e(A,B)$$

et

$$e(A,B) = e(-A, -B) \qquad e(A, -B) = e(-A,B).$$

Théorème III : *Ecarts relatifs et distance relative sont équivalents.*

$$|A| \geq |B| \Longrightarrow \frac{e(A,B)}{d(A,B)} = |a| \Longrightarrow$$

$$\mathcal{B}^{-1} \leq \frac{e(A,B)}{d(A,B)} = |a| < 1$$

sinon l'on permute A et B donc a et b

De même, les écarts relatifs définis sur deux bases différentes sont équivalents.

$$A = a_1 \mathcal{B}_1^{\alpha_1} = a_2 \mathcal{B}_2^{\alpha_2} \qquad \mathcal{B}_1^{\alpha_1} : \mathcal{B}_2^{\alpha_2} = a_2 : a_1$$

$$\frac{e_1(A,B)}{e_2(A,B)} = \frac{|A-B|}{\mathcal{B}_1^{\max(\alpha_1,\beta_1)}} : \frac{|A-B|}{\mathcal{B}_2^{\max(\alpha_2,\beta_2)}} = \frac{\mathcal{B}_2^{\alpha_2}}{\mathcal{B}_1^{\beta_1}} = \frac{|a_1|}{|a_2|} \text{ (si } |A| \geq |B|)$$

$$\mathcal{B}_1^{-1} < \frac{e_1(A,B)}{e_2(A,B)} < \mathcal{B}_2$$

Mentionnons aussi les théorèmes suivants :

1/ $0 \leq e(A,B) < 2$

2/ $\mathcal{B}^{-1} \leq e(A,0) < 1 \qquad \forall A \neq 0$

3/ $e(|A|,|B|) = \begin{cases} e(A,B) & AB > 0 \\ e(A,-B) & AB < 0 \end{cases}$

4/ $e(|A|,|B|) \leq e(A,B) \leq e(|A|,-|B|)$

$e(|A|,|B|) \leq e(A,-B) \leq e(|A|,-|B|)$

Ceci se démontre en remarquant que :

$$||A| - |B|| \leq |A + B| \leq |A| + |B|$$

5/ $AB \geq 0 \Longrightarrow e(A,B) < 1$

$AB < 0 \Longrightarrow e(A,B) \geq \mathcal{B}^{-1}$

$1 < e(A,B) < 2 \Longrightarrow AB < 0$

$0 < e(A,B) < \mathcal{B}^{-1} \Longrightarrow AB > 0$

6/ $\alpha = \beta$ $\begin{cases} (AB \geq 0) \Longrightarrow 0 \leq e(A,B) < 1 - B^{-1} \\ (AB \quad 0) \Longrightarrow 2B^{-1} \leq e(A,B) < 2 \end{cases}$

en effet : $\left.\begin{matrix} B^{-1} \leq |a| < 1 \\ B^{-1} \leq |b| < 1 \end{matrix}\right\} \Longrightarrow \begin{cases} ||a|-|b|| < 1 - B^{-1} \\ |a|+|b| \geq 2B^{-1} \end{cases}$

7/ $e(A,B) < B^{-1} - B^{-2} \Longrightarrow |\alpha - \beta| < 2$ et $AB \geq 0$

en effet : $|\alpha-\beta| < 2 \Longrightarrow 0 \leq e(A,B) < 1 - B^{-2}$

$|\alpha-\beta| \geq 2 \Longrightarrow B^{-1} - B^{-2} < e(A,B) < 1$

$e(A,B) > 1+B^{-1} \Longrightarrow \alpha=\beta$ et $AB < 0$

en effet : $|\alpha-\beta| \geq 1 \Longrightarrow B^{-1} \leq e(A,B) < 1+B^{-1}$

Il peut être intéressant de noter, en binaire :

$e(A,B) < \frac{1}{4} \Longrightarrow \frac{1}{2} < \frac{A}{B} < 4 \quad (|A| \geq |B|)$

A et B ont le même exposant à "1 près".

NOTATIONS :

Pour les figures (2a) et (2b) nous avons utilisé les notations suivantes :

$B^{-1}$ : $(A=0)U(B=0) \Longrightarrow$ : 1

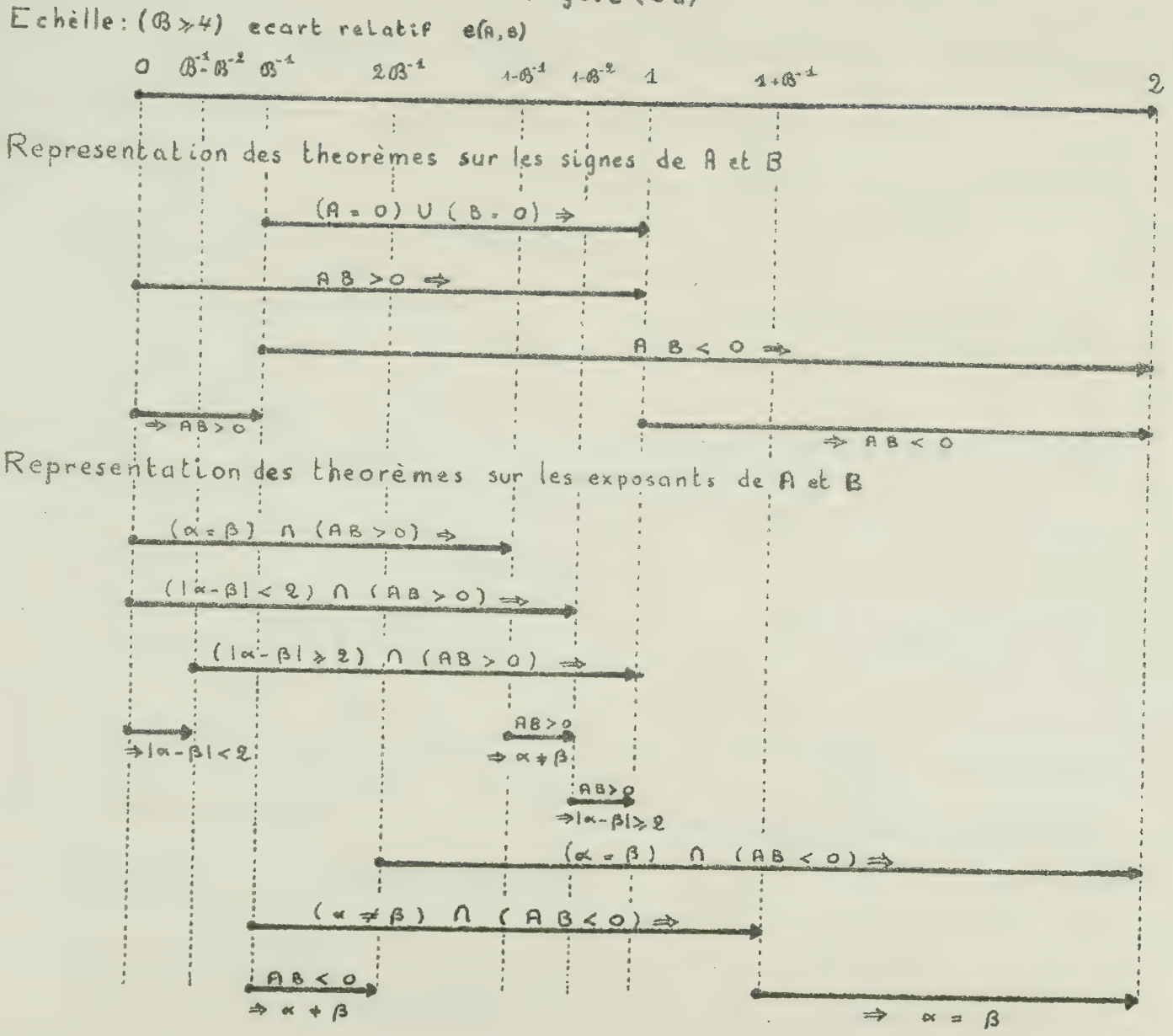
signifie $\left.\begin{matrix} A=0 \\ \text{ou } B=0 \end{matrix}\right\} \Longrightarrow B^{-1} \leq e(A,B) < 1$

$B^{-1}$ : $AB < 0$ : $2B^{-1}$

$\Longrightarrow \alpha \neq \beta$

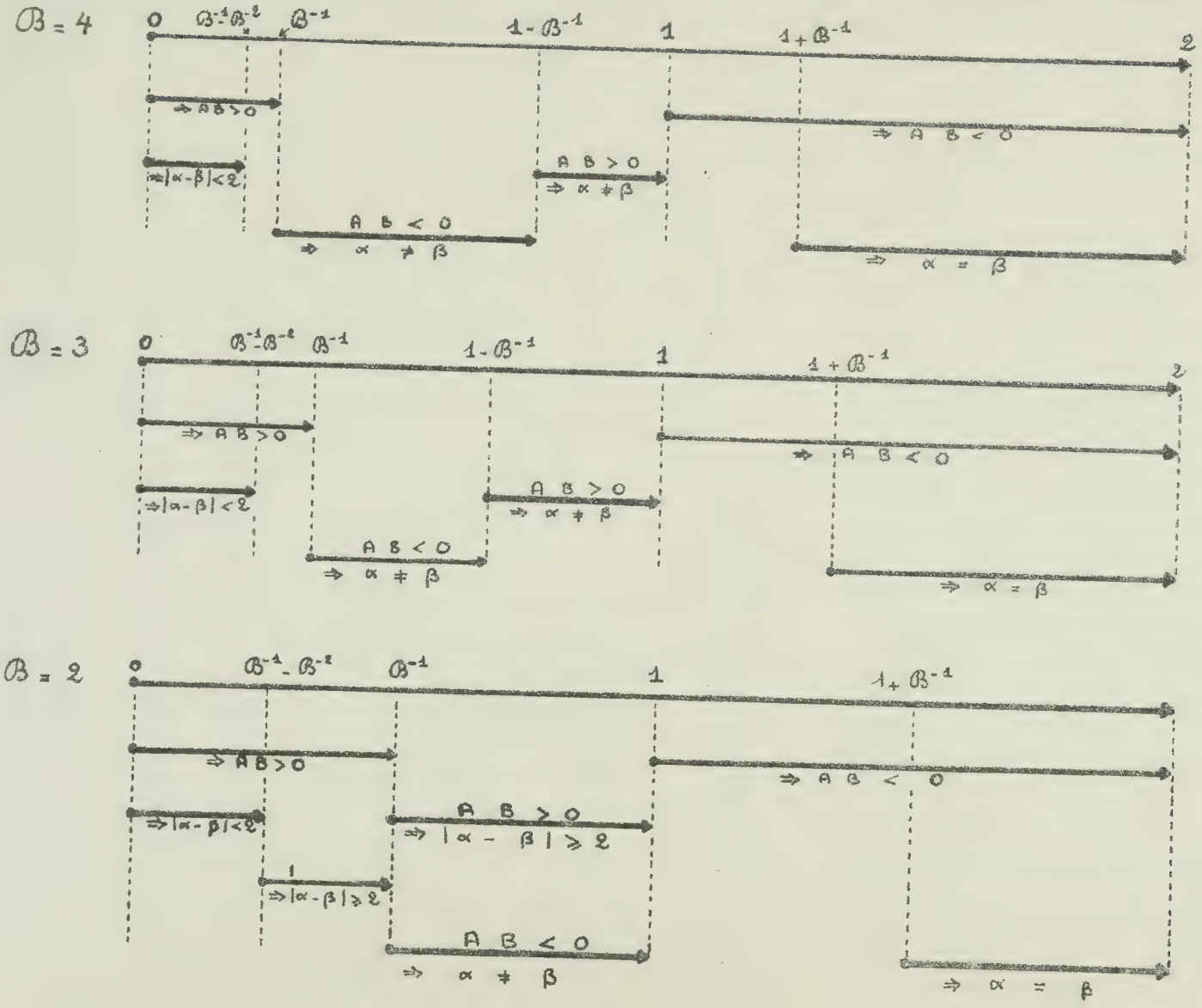
signifie $AB<0$ et $B^{-1} \leq e(A,B) < 2B^{-1}$ $\Longrightarrow \alpha \neq \beta$

Figure (2 a)


Echelle: (B ≥ 4) ecart relatif e(A,B)
0
B⁻¹-B⁻²
B⁻¹
2B⁻¹
1-B⁻¹
1-B⁻²
1
1+B⁻¹
2
Representation des theorèmes sur les signes de A et B
(A = 0) ∪ (B = 0) ⇒
AB > 0 ⇒
A B < 0 ⇒
⇒ AB > 0
⇒ A B < 0
Representation des theorèmes sur les exposants de A et B
(α = β) ∩ (AB > 0) ⇒
(|α-β| < 2) ∩ (AB > 0) ⇒
(|α-β| ≥ 2) ∩ (AB > 0) ⇒
AB > 0
⇒ |α-β| < 2
AB > 0
⇒ α ≠ β
AB > 0
⇒ |α-β| ≥ 2
(α = β) ∩ (AB < 0) ⇒
(α ≠ β) ∩ (A B < 0) ⇒
AB < 0
⇒ α ≠ β
⇒ α = β

# Figure (2 b)

Representation des theorèmes sur l'ecart relatif $e(A,B)$ pour des bases particulières.



## I.7. SIGNIFICATION DE L'ADDITION ET DE LA SOUSTRACTION DE DEUX NOMBRES-MACHINE.

Dans le cas particulier de l'addition (ou soustraction) de deux nombres-machine en virgule flottante normalisée, seule considérée ici, *remarquons que* e(A,-B) ou e(A,B) *représente ce que tout ordinateur est nécessairement amené à calculer pour réaliser ces opérations.* Notons encore que la soustraction se ramène à une addition en changeant le signe du second opérande, transformant ainsi une opération disymétrique en A et B (A-B ≠ B-A) en une opération symétrique. *Nous pouvons alors supposer que le premier opérande de l'addition est le plus grand en module.*

Le calcul de la mantisse-résultat se décompose en quatre traitements principaux :

| | | |
|---|---|---|
| . Cadrage des mantisses | 11100110 | a |
| | 11001111 | $b\mathcal{B}^{\beta-\alpha}$ |
| . Addition des mantisses cadrées | 10001100111 | $a+b\mathcal{B}^{\beta-\alpha}$ |
| . Troncature (avec ou sans arrondi) | 100011010 | |
| . Normalisation éventuelle | 10001101 | |

1/ Le cadrage des mantisses introduit une perte d'information par non-utilisation des chiffres de plus faible poids du plus petit nombre.

Cette perte est nulle lorsque les exposants sont égaux :

$$1 + \mathcal{B}^{-1} < e(|A|,-|B|) < 2 \Longrightarrow \alpha=\beta \Longrightarrow \text{perte nulle}$$

et positive dans tous les autres cas, notamment lorsque $e(|A|,-|B|) > 2\mathcal{B}^{-2}$ , seul cas où l'on puisse affirmer la non égalité des exposants.

Elle est maximale lorsque le quotient des 2 nombres est inférieur à $B^{-n}$. Lors du cadrage, le nombre le plus petit est alors indiscernable de zéro, relativement au plus grand : on dit qu'il est (relativement) négligeable :

B négligeable devant A $\Longleftrightarrow (|B/A| < B^{-n})$

$\Longleftrightarrow e(A,B) = |a| \Longleftrightarrow e(A,A+B) = e(A,A-B) = 0$

2/ 3/ L'addition des mantisses cadrées est réalisée par addition arithmétique sur les chiffres de gauche (significatifs). Un arrondi est fait sur le premier chiffre négligé du plus petit opérande, selon la convention choisie par la troncature.

4/ A - Lorsque l'additon des mantisses donne un résultat supérieur à 1, il est nécessaire de normaliser, ce qui entraîne la perte d'un chiffre à la suite du décalage à droite. Pour une addition :

Perte d'information par dépassement de capacité $\Longleftrightarrow 1 \leq e(A,-B) < 2$

Signalons l'existence d'un autre dépassement de capacité, obtenu lorsque l'ordre de grandeur du nombre-machine est trop grand ou trop petit. Certains ordinateurs indiqueront que l'opération demandée est impossible à réaliser. En CAB 500, le nombre-résultat est pris égal au plus grand nombre représentable (en module) ou à zéro, mais cette convention se révèle à l'usage d'un emploi délicat.

4/ B - Lorsque le résultat de la soustraction de deux nombres-machine de même signe comporte des zéros principaux ($|a-b\ B^{\beta-\alpha}| < B^{-1}$), la normalisation introduit les chiffres non significatifs (ou chiffres de bruit). Nous les notons (?) signalant ainsi que ces chiffres ont une valeur inconnue.

$e(A,B) < B^{-l} \Longrightarrow$ Chiffres de bruit

Par exemple, en convention décimale :

| | | |
|---|---|---|
| A | : 0.77 563 $10^9$ | |
| B | : 0.77 514 $10^9$ | |
| A-B | : 0.00 049 $10^9$ | avant normalisation |
| A-B | : 0.49 000 $10^6$ | après normalisation |

En pratique, on donne aux chiffres de bruit la valeur zéro et il apparaîtra alors le résultat 0,49 000 $10^6$.

Remarquons cependant que la précision de la mantisse-résultat n'est plus $B^{-n}$ (dans l'exemple : $10^{-5}$) où n est le nombre de chiffres de la mantisse mais $B^{-m}$ (ici : $10^{-2}$) où m indique le nombre de chiffres significatifs.

$e(A,B) < B^{-k} \Longrightarrow$ au moins k chiffres de bruit.

Lorsqu'un ordinateur effectue une soustraction de deux nombres ayant la même représentation :

$A^* = B^* = \{a,\alpha\}$ , $e(A,B) = 0$ et $|a-b| < B^{-n}$

le résultat vérifie la relation :

$$0 \leq |A \ominus B| \leq B^{-n}$$

Ce que nous représenterons symboliquement par :

$$A^* \ominus B^* = \{0, \alpha - n\}$$

en donnant un sens à l'exposant de zéro.
Le nombre $Z = \{0,\zeta\}$ (dit "zéro relatif") représente alors tout réel négligeable devant $A = \{a,\zeta + n\}$

$$\forall a, \quad B^{-l} \leq |a| < 1 \quad ; \quad A \oplus Z = A \ominus Z = A$$

La définition et l'utilisation du zéro relatif en ordinateur nécéssitent une modification importante des microprogrammes réalisant les opérations élémentaires.

Plus généralement, la création de bruit, non signalée par l'arithmétique de virgule flottante normalisée, conduit à concevoir de nouvelles arithmétiques actuellement à l'étude au Laboratoire de Besançon et que nous indiquerons à la fin de ce chapitre.

## I.8. REMARQUES SUR LA MULTIPLICATION, LE PRODUIT SCALAIRE ET LA DIVISION.

La multiplication en virgule flottante se ramène à un double traitement: addition des exposants, multiplication des mantisses. Il y a lieu de prévoir aussi une normalisation éventuelle du nombre-résultat.

. L'addition des exposants ne pose aucun problème particulier puisqu'il s'agit d'une opération portant sur des nombres entiers. Il peut certes y avoir dépassement de capacité ($\alpha+\beta$ ne pouvant être représenté) mais nous n'approfondirons pas ici cette question à l'exception d'une remarque qui lui sera consacrée.

. La multiplication des mantisses peut s'opérer avec ou sans arrondi (ou encore avec arrondi aléatoire). Dans le premier cas, le plus généralement adopté pour des raisons d'efficacité, remarquons qu'il y a nécessité de calculer au moins $(n+1)$ chiffres significatifs, ($n$ étant le nombre commun de chiffres définissant la mantisse des opérandes et du résultat).

. $B^{-1} \leq |a| < 1$, $B^{-1} \leq |b| < 1 \Longrightarrow B^{-2} \leq |axb| < 1$

Il peut donc y avoir création d'un zéro à gauche (zéro principal). L'ordinateur normalise alors, en raison même du programme qui lui aura été imposé et de la nécessité de traiter des nombres normalisés.

Par exemple : A = {0.29863,3} B = {0.16728,5}

A x B = {0.0499548264,8} → {0.49955,7}

L'opération symbolisée par la flèche (→) n'est possible que si l'on a conservé les "chiffres de garde" 4(premier chiffre de "droite") et 8(à cause de l'arrondi). Il y a donc lieu de programmer la multiplication des mantisses de manière à s'assurer d'au moins (n+2) chiffres dans le produit avant normalisation.

Une opération est fondamentale, notamment en calcul matriciel, (mais sa signification dépasse de beaucoup ce cas particulier) : LE PRODUIT SCALAIRE que l'on peut noter :

$$S(A,B) = \sum_{i=1}^{n} A[i] \times B[i]$$

Le calculer par le programme suivant :

$$S(A,B) = (A[1] \times B[1]) + (A[2] \times B[2]) + \ldots\ldots (A[N] \times B[N])$$

conduirait à un résultat porteur d'une erreur somme des erreurs partielles commises lors de chaque arrondi, soit N0 par exemple. Il est donc nécessaire d'effectuer chaque multiplication $A[I] \times B[I]$ en conservant au moins $(n+2+\log_B N)$ chiffres dont $(2 + \log_B N)$ chiffres de garde et chaque addition sur la totalité des chiffres significatifs.

Dans l'ignorance de la valeur N, il est recommandé de

remplacer $(n+2+\log_\beta N)$ par $2n$(double précision), qui est en général une majoration et même une forte majoration ($N << \beta^{n-2}$).

Lors d'une multiplication de deux nombres-machine très grands en valeurs absolues, le résultat peut être trop grand et dépasser l'infini-machine $\infty^*$. Cette difficulté, déjà signalée pour l'addition peut être levée de deux manières : soit signaler le dépassement de capacité (le résultat sera alors l'information "Trop Grand") soit prendre comme résultat le plus grand nombre représentable (en module). Cette conception, mise en oeuvre sur CAB 500, présente en fait de sérieux inconvénients (toute opération sur $\infty^*$ ne donne pas $\infty^*$ comme résultat). Il peut donc être souhaitable de traiter $\infty^*$ comme un nombre spécial (au même titre que le zéro) ; alors toute multiplication par $\infty^*$ donnerait $\infty^*$. Le problème $\infty^* \times 0$ reste critique.

Comme la multiplication, la division en virgule flottante se ramène à un double traitement : soustraction des exposants, division des mantisses :

$\beta^{-1} \leq |a| < 1$, $\beta^{-1} \leq |b| < 1 \Longrightarrow \beta^{-1} < |a/b| < \beta$

Trois cas sont possibles

. $\beta^{-1} < |a/b| < 1$ : aucune normalisation n'est nécessaire

. $|a/b| = 1$ : un test spécial fonctionne alors

. $1 \leq |a/b| < \beta$ : un décalage d'un chiffre vers la droite est nécessaire. On voit que la normalisation de la représentation de (A/B) conduit aux mêmes problèmes que pour la multiplication : il importe de conserver au moins deux chiffres de garde dans le calcul du quotient. La division par zéro arrête généralement le fonctionnement de l'ordinateur.

## I.8. PROGRAMME DE TRONCATURE OU SIMULATEUR DE PRECISION VARIABLE -

Les estimateurs d'erreurs proposées par certains auteurs (en particulier Monsieur CASTINEL [2] ou Monsieur MAITRE [3] ) malgré leur grand intérêt théorique, ne serrent pas toujours suffisamment la réalité.

Dans ces conditions, il serait souhaitable que l'ordinateur indique automatiquement une marge d'erreur ou signale l'imprécision des résultats. Une des méthodes possibles est celle de la "troncature" : on effectue deux fois le même calcul avec des précisions relatives différentes et la comparaison des résultats donnera alors une idée de l'erreur possible supposée linéaire. Comme la précision d'un nombre écrit en virgule flottante dépend essentiellement du nombre n des chiffres de la mantisse, changer la précision relative revient à modifier ce nombre n. Il est souvent prévu sur un ordinateur, des nombres en virgule flottante normalisée dit en "simple" ou "double longueur" et cette possibilité est mise à profit par les compilateurs ALGOL qui prévoieront plusieurs type de déclaration de Réel. Mais cette modification de la précision nécessite de transformer profondément tout programme langage-machine ou de changer toutes les déclarations de Réel d'un programme ALGOL. Au contraire, nous avons introduit ce changement d'arithmétique au niveau même des microprogrammes CAB 500, rendant le choix de la précision indépendant de l'algorithme. Le simulateur d'arithmétique variable, que nous avons spécialement écrit pour CAB 500, permet de faire varier avec souplesse le nombre n de chiffres de la mantisse pour obtenir ainsi différentes précisions, inférieures ou égales à celle fournie par le constructeur.

Pour cela, il a été nécessaire d'étudier le fonctionnement de chaque microprogramme utilisé en calcul matriciel (addition et soustraction, multiplication, division, racine carrée, lecture et impression). Cette étude nous a été facilitée par l'ouvrage de Linski [4] et par les quelques notes fournies par le constructeur [5] .

Il a été rédigé un programme de troncature commun vers lequel s'aiguille le microprogramme lorsque le traitement qui lui est propre est terminé.

Il y a donc seulement changement de précision, le traitement propre à chaque microprogramme restant inchangé. Toutes les opérations sont simulées d'une manière identique et seul le résultat est tronqué du nombre de chiffres désirés. En particulier, il a été prévu deux utilisations possibles avec ou sans lecture tronquée, pour éprouver l'influence de la définition initiale des nombres sur le déroulement des calculs. L'impression, considérée comme une simple traduction de convention binaire en convention décimale n'est jamais simulée.

Ce programme de troncature se compose de deux parties :

a/ Une partie "exécution" que nous avons cherché à rendre rapide et sûre. Elle teste tous les débordements et fait les rétablissements nécessaires. La CAB 500 étant un ordinateur à programme enregistré sur tambour, nous avons spécialement étudié le cas où aucune simulation n'est demandée. La précision et la vitesse de calcul sont alors les mêmes que celles indiquées par le constructeur (précision maximale), et pour éviter que des utilisateurs ne fassent involontairement des calculs dans une précision moindre, l'ordinateur se met automatiquement en précision maximale à toute nouvelle entrée de programme.

b/ Une partie "commande" où se fait le choix de l'arithmétique. Il est possible d'effectuer la troncature de trois façons :

1) troncature de l chiffres binaires ($0 \leq \ell \leq 24$)

2) troncature avec arrondi préalable

3) suivant le résultat d'un test interne, cet arrondi sera ou non fait avant la troncature (arrondi aléatoire).

Pour préciser l'état du simulateur, il est donc nécessaire d'indiquer trois paramètres, et cela peut se faire à tout moment, avant ou pendant le déroulement des calculs.

1er paramètre : indique le nombre de chiffres à tronquer

2e paramètre : indique si l'on veut ou non un arrondi

3e paramètre : indique si l'on veut faire l'arrondi suivant un choix aléatoire et nous noterons l'état du simulateur par $(\ell)$, $(\ell A)$, $(\ell \wedge A)$, $\ell$ étant un nombre entier inférieur à 24.

Ces notations seront utilisées dans les tableaux où nous résumons les essais effectués :

Pour l'utilisation de ce programme de troncature à l'intérieur d'un programme ALGOL (changement d'arithmétique en cours de calcul) les trois paramètres seront une variable entière et deux variables booléennes de la procédure "simulation"

ainsi définie :

Procédure Simulation (l, arrondi, aléatoire) ;
Valeur l ; Entier l ; Booleen arrondi, aléatoire ;
Début

Corps en langage machine

Fin ;

Nous avons étudié cette procédure sur CAB 500, mais comme le compilateur ALGOL ne prévoit pas les procédures à corps en langage machine, la déclaration de cette procédure est alors très délicate et doit être confiée à un spécialiste connaissant bien l'ordinateur et son compilateur.

Essais effectués :

Nous avons utilisé ce simulateur pour le calcul de fonctions trigonométriques, considérées comme des polynômes de degré peu élevé ayant des coefficients aux propriétés particulières (alternés, décroissants en module) et pour l'étude de quelques polynômes correspondant aux matrices tests indiquées au chapitre IV. D'autres utilisations assez particulières, ne se rattachant pas à cette étude.

## CONCLUSIONS DU CHAPITRE I

Nous avons voulu présenter ici les nombres réels tels qu'ils sont représentés en virgule flottante normalisée dans un ordinateur, mettant en évidence le traitement particulier qu'ils subissent dans les opérations élémentaires et leurs différences profondes avec les nombres réels définis et utilisés en mathématiques.

Nous avons vu la nécessité de définir sur l'ensemble des réels-machine une pseudo-distance, l'écart relatif, mieux adapté à la mesure de ses éléments que des distances (telles que la distance relative) définie sur l'ensemble des réels. L'écart relatif se révèle intéressant pour suivre le traitement de l'information et s'intègre facilement à l'ensemble des microprogrammes puisque tout ordinateur est amené à le calculer lors du microprogramme addition-soustraction.

Pour ces nombres-machines nous avons signalé plusieurs estimations d'erreur. Celle de Monsieur CASTINEL très proche du langage mathématique, met en évidence la précision relative de tels nombres et a été critiquée. Nous avons proposé et essayé une "estimation automatique" d'erreur en étudiant le programme de troncature, constatant que, malgré les avantages de cette simulation, elle reste rarement utilisée parce qu'elle est lourde d'emploi et ne détecte pas (non plus) la création de bruit.

Il nous semble alors intéressant de rechercher une "indication automatique" d'une marge d'erreur dans deux autres directions. Un simulateur d'arithmétique opérerait sur des

nombres en virgule flottante où serait indiqué le nombre de chiffres encore significatifs. Parmi plusieurs méthodes possibles, on pourrait choisir de représenter un nombre réel par trois paramètres, la mantisse normalisée et l'exposant comme dans la convention traditionnelle et un compteur de chiffres significatifs. Toute perte de précision (ou bruit) se traduirait par une diminution de ce nombre. Cette nouvelle arithmétique ressemble globalement à celles proposées par Monsieur METROPOLIS [6].

La deuxième direction de recherche est toute différente car il ne s'agit plus de modifier l'arithmétique au niveau élémentaire mais de transformer profondément la structure de virgule flottante lors de calcul matriciel. En effet, lorsque nous parlons de nombres petits ou grands, nous sous-entendons une comparaison qu'il est délicat de préciser. Il est par contre souhaitable en calcul matriciel, de considérer que les éléments de la matrice forment un ensemble et un nombre "petit" le sera par rapport à l'élément maximum (en module). De la même manière, la précision relative d'un élément sera définie par rapport à l'élément maximum et non par rapport à lui-même. Cette arithmétique nécessitant des opérations élémentaires propres au calcul matriciel doit être intégrée à un ensemble capable seulement de faire des calculs matriciels. Elle est très voisine de celle proposée par WILKINSON sous le nom de floating block [7].

Ces deux types d'arithmétiques actuellement à l'étude au Laboratoire de Besançon, feront l'objet d'un mémoire ultérieur [8].

# TRONCATURE (Chargement)

******

| | | |
|---|---|---|
| 91,0 | A12EV12163 | CO RETOUR = 95,8. |
| 91,1 | EV25 | |
| 91,2 | EV51 | S |
| 91,3 | EV41 | I |
| 91,4 | EV45 | M |
| 91,5 | EV53 | U |
| 91,6 | EV44 | L |
| 91,7 | EV15 | |
| 91,8 | A9EL0 | 1er Paramètre N < 18 |
| 91,9 | T9R10 | |
| 91,10 | S10V18 | |
| 91,11 | RV91,1P10 | |
| 91,12 | A10EL0 | 2ème Parametre (A ou RC) |
| 91,13 | V10TV25 | |
| 91,14 | RV91,25Z13 | |
| 91,15 | V10TV33 | |
| 91,16 | RV91,1Y13 | |
| 91,17 | A11EL0 | 3ème Parametre (A ou RC) |
| 91,18 | V11TV25 | |
| 91,19 | RV9125Z13 | |
| 91,20 | V11TV33 | |
| 91,21 | RV91,1Y13 | |
| 91,22 | A14EL0 | Retour . Chariot ? |
| 91,23 | V14TV25 | |
| 91,24 | RV91,1Y13 | |
| 91,25 | RV91,29Y9 | Troncature . |
| 91,26 | TL2R13 | Pas de Troncature |
| 91,27 | RCV95,120 | |
| 91,28 | RV91,74 | |
| 91,29 | V10TV25 | |
| 91,30 | RV91,36Y13 | |
| 91,31 | A14EM91,32S | Cas SIMUL O |

TRONCATURE (Chargement)

******

| | | |
|---|---|---|
| 91,0 | | |
| 91,32 | RV95,76 | ORDRE MIS EN MEMOIRE |
| 91,31 → 91,33 | T14EM127,71 | DANS 127,71 |
| 91,34 | T14M127,70 | INEFFECTIF DANS 127,70 |
| 91,35 | RV91,56 | |
| 91,30 → 91,36 | V11TV25 | |
| 91,37 | RV91,43Y13 | |
| 91,38 | A14EM91,398 | |
| 91,39 | RV95,73 | ORDRE EN MEMOIRE |
| 91,40 | T14EM127,71 | DANS 127,71 |
| 91,41 | T14M127,70 | INEFFECTIF DANS 127,70 |
| 91,42 | RV91,52 | |
| 91,43 | A14EM91,448 | |
| 91,44 | I10TW128 | ORDRE EN MEMOIRE |
| 91,45 | T14EM127,70 | DANS 127,70 |
| 91,46 | A14EM91,478 | |
| 91,47 | RV95,73Z13 | ORDRE EN MEMOIRE |
| 91,48 | T14EM127,71 | DANS 127,71 |
| 91,49 | A14EM91,508 | |
| 91,50 | RV95,76 | ORDRE EN MEMOIRE |
| 91,51 | T14EM127,72 | DANS 127,72 |
| 91,52 | T15R10 | |
| 91,53 | T9R15 | |
| 91,54 | A14EM91,81+ | PUISSANCE DE 2 |
| 91,55 | T10R15 | |
| 91,56 | T14EM127,74 | Pour arrondi |
| 91,57 | A14V64 | |
| 91,58 | K14TV1 | CREATION |
| 91,59 | T13R14 | DU MASQUE |
| 91,60 | S9V1 | DE |
| 91,61 | RV91,57Y9 | TRONCATURE |
| 91,62 | T14EM127,77 | DANS 127,77. |
| 91,63 | X | |

TRONCATURE (Chargement)



******

| | | | |
|---|---|---|---|
| 91,0 | 91,64 | K | |
| | 91,65 | K | |
| | 91,66 | RV91,70 | |
| | 91,67 | ~~T10M127,68~~ | } Exterieur au PROGRAMME. |
| | 91,68 | K | |
| | 91,69 | ~~RV96,6~~ | |
| | 91,70 | K | |
| | 91,71 | A14EM91,72S | |
| | 91,72 | RV95,64 | ORDRE EN MEMOIRE |
| | 91,73 | T14EM127,63 | DANS 127,63. |
| 91,28 | 91,74 | RETOUR | |

******

******

TRONCATURE (Execution)

Annexe

1.31.

******

| | | | |
|---|---|---|---|
| 95,56 | 95,64 | T15ER9 | RANGEMENT R15. |
| | 95,65 | A15ER11 | |
| | 95,66 | C15QV27 | |
| | 95,67 | A10ER0+ | Appel OPERANDE |
| | 95,68 | A12EV0 | |
| | 95,69 | RV127,70 | |
| | 95,70 | X | sans/avec Arrondi |
| | 95,71 | X | |
| | 95,72 | X | |
| | 95,73 | A10M127,74S | Addition Arrondi |
| | 95,74 | X | |
| | 95,75 | X | |
| 91,31 → | 95,76 | I10QM127,77S | Masque de TRONCATURE |
| | 95,77 | X | |
| | 95,78 | S10R14 | TRONCATURE |
| ↓ | 95,79 | RV95,99B10 | Pas de Correction |
| | 95,80 | I10QV127 | Exposant |
| | 95,81 | K10TV25 | |
| | 95,82 | A13W32768D13 | |
| | 95,83 | S10R14 | Mantisse seule. |
| | 95,84 | C10QV1 | |
| | 95,85 | A13W256 | Correction |
| | 95,86 | RV95,94B13 | Mantisse. |
| | 95,87 | RV95,92N13 | |
| | 95,88 | A10EW49152N10 | |
| | 95,89 | A10EW16384P10 | |
| | 95,90 | S10V128P10 | |
| | 95,91 | A10V63 | |
| | 95,92 | A10EVON13 | |
| | 95,93 | RV95,99 | |
| | 95,94 | C13QV25 | |
| ↓ | 95,95 | I13QV127 | |

TRONCATURE (Execution)

*******

| | | | |
|---|---|---|---|
| 95,56 | 95,96 | I1OTV127 | |
| | 95,97 | S1OR13 | |
| | 95,98 | A1OR14 | |
| 95,79 } | 95,99 | RR12Y12 | Retour à troncation lecture. |
| 95,93 } | 95,100 | A15ER11 | |
| | 95,101 | C15TV27 | |
| | 95,102 | T1ORO+ | |
| | 95,103 | T9ER15 | |
| | 95,104 | UR11 | Sortie avec Rangement Registre. |
| | 95,105 | X | |
| | 95,106 | X | |
| | 95,107 | X | |
| | 95,108 | X | |
| | 95,109 | X | |
| | 95,110 | X | |
| | 95,111 | X | |
| | 95,112 | X | |
| | 95,113 | X | |
| | 95,114 | X | |
| | 95,115 | X | |
| | 95,116 | X | |
| | 95,117 | X | |
| | 95,118 | X | PAS DE TRONCATURE (SIMUL O) |
| 95,1 → | 95,119 | A12EV12163 | (R12) = 95,8 |
| 95,27 → | 95,120 | A14EM95,121S | |
| | 95,121 | UR11 | ORDRE EN MEMOIRE |
| | 95,122 | T14EM127,63 | DANS 127,63. |
| | 95,123 | A14EM95,124S | |
| | 95,124 | RR12 | ORDRE EN MEMOIRE |
| | 95,125 | T14EM127,70 | DANS 127,70 |
| | 95,126 | G12TVO | PERMUTER R12/R13. |
| | 95,127 | RR13 | |

*******

*******

TRONCATURE (EXECUTION)

****** TRONCATION DE LA LECTURE

```
95,56     9556     X
          95,57    X
          95,58    X
          95,59    T12R10
          95,60    RCV127,70    TRONCATION
          95,61    T10R12
          95,62    RV90,103D11  RANGEMENT  MAGNETIQUE
          95,63    RV90,111     RANGEMENT  REGISTRE
```

TRONCATURE (Modification Microprogrammes).

*******

| | | | |
|---|---|---|---|
| 95,1 | 95,1 | RV95,119 | {REMISE A ZERO<br>SIMUL |

*******

*******

| | | | |
|---|---|---|---|
| 110,62 | 110,62<br>110,63 | RV127,63213<br>RR11 | VERS ENTREE SIMUL |

*******

*******

| | | | |
|---|---|---|---|
| 90,107 | 90,107 | RV95,59 | TRONCATURE LECTURE. |

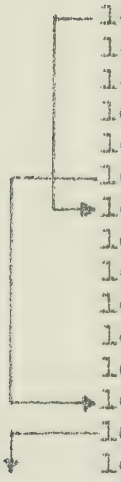
*******

*******

| EXECUTION | ORDRE | CHARGEMENT | SIGNIFICATION |
|---|---|---|---|
| 127,63 | UR11 | 95,122 | PAS DE TRONCATURE |
| | RV95,64 | 91,73 | TRONCATURE |
| 127,70 | RR12 | 95,125 | PAS DE TRONCATURE |
| | X | 91,34 | TRONCATURE SANS ARRONDI |
| | X | 91,41 | AVEC ARRONDI |
| | ILOTW128 | 91,45 | AVEC ARRONDI ALEATOIRE |
| 127,71 | RV95,76 | 91,33 | SANS ARRONDI |
| | RV95,73 | 91,40 | AVEC ARRONDI |
| | RV95,73Z13 | 91,47 | AVEC ARRONDI ALEATOIRE |
| 127,72 | RV95,76 | 91,51 | AVEC ARRONDI ALEATOIRE |
| 127,74 | PUISSANCE de 2 | 91,56 | AVEC ARRONDI<br>AVEC ARRONDI ALEATOIRE |
| 127,77 | MASQUE de TRONCATURE | 91,62 | TRONCATION. |

ECART RELATIF

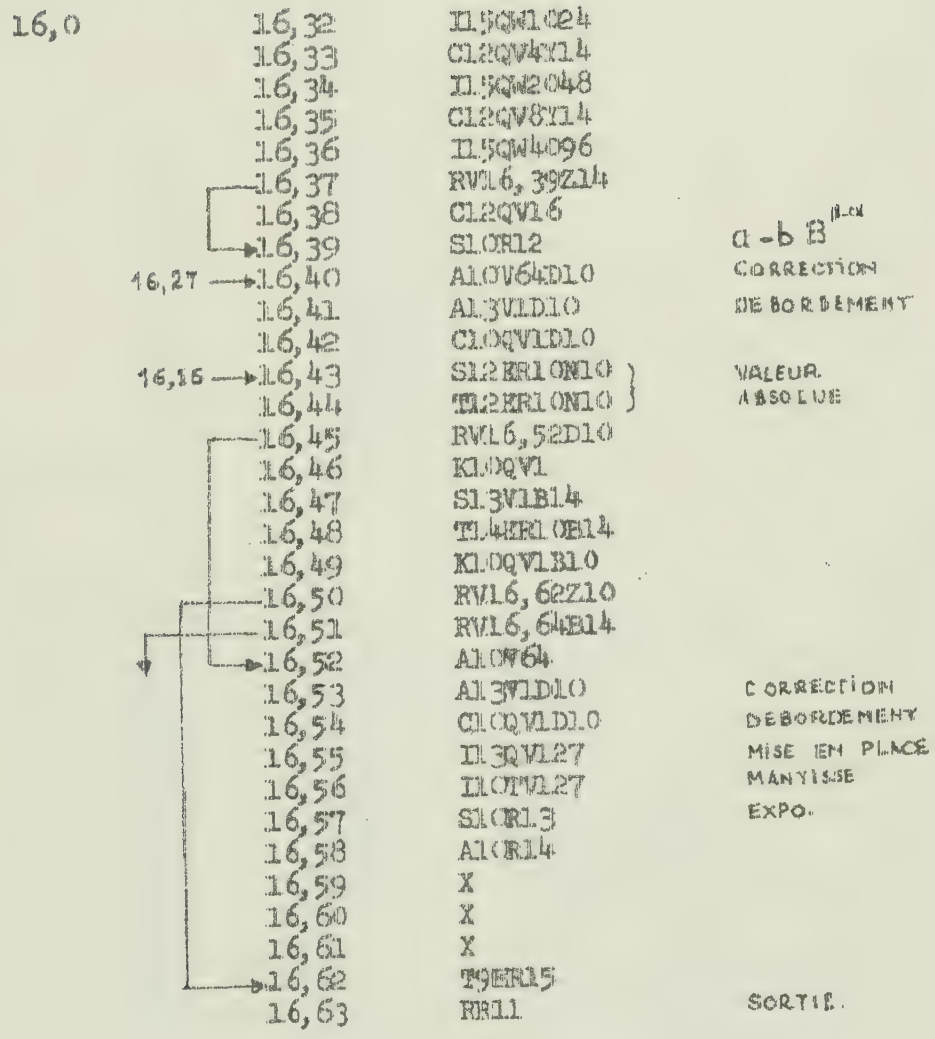
*******



| | | |
|---|---|---|
| 16,0 | T12EE11 | Rangement CO de retour |
| 16,1 | T9ER12 | 2ème OPERANDE B |
| 16,2 | T15ER9 | Rangement du Registre 15 |
| 16,3 | RV16,9Z10 | 1er OPERANDE A NUL |
| 16,4 | I10QY127 | Exposant $\alpha$ |
| 16,5 | K10Y25 | |
| 16,6 | S10R14 | mantisse $a$ |
| 16,7 | A13W32768D13 | |
| 16,8 | RV16,15Z12 | B = 0 |
| 16,9 | I12QY127 | Exposant $\beta$ |
| 16,10 | T14ER15 | |
| 16,11 | K12QY25 | |
| 16,12 | S12R15 | mantisse $b$ |
| 16,13 | A14W32768D14 | |
| 16,14 | T12ER10Z10 | |
| 16,15 | T12R13Z12 | |
| 16,16 | RV16,43Z12 | un des OPERANDES est nul |
| 16,17 | X | |
| 16,18 | X | |
| 16,19 | S14R13 | $\beta - \alpha$ |
| 16,20 | A15ER12P14 | Permutation |
| 16,21 | A12ER10P14 | A = le plus grand en module |
| 16,22 | A10ER15P14 | B = " petit " |
| 16,23 | A13EY0 | |
| 16,24 | A15ER14P14 | $\lvert\alpha - \beta\rvert$ |
| 16,25 | S15ER14N14 | |
| 16,26 | I15QV24576 | |
| 16,27 | RV16,40Y14 | $\lvert\alpha - \beta\rvert > 63$ |
| 16,28 | I15QV256 | Decalage |
| 16,29 | C12QY1Y14 | du |
| 16,30 | I15QY512 | second |
| 16,31 | C12QY2Y14 | OPERANDE. |

ECART RELATIF

******



~~*****~~

| | | | |
|---|---|---|---|
| ~~16,0~~ | 16,64 | C1 0QV8 | Exposant négatif. |
| | 16,65 | S1 3V22 | |
| | 16,66 | D1 0QV1 | |
| | 16,67 | A1 3V1Y1 0 | |
| | 16,68 | RV1 6,66Y1 0 | |
| | 16,69 | A1 0ZR14 | |
| | 16,70 | V1 0QW57 344 | |
| | 16,71 | EV2 3Z14 | DETECTION |
| | 16,72 | EV3 9Z14 | ERREUR |
| | 16,73 | EV2 4Z14 | |
| | 16,74 | FZ14 | |
| | 16,75 | RV1 6,52 | |
| | 16,76 | X | |
| | 16,77 | X | |
| | 16,78 | X | |
| | 16,79 | X | Exemple d'utilisation en PAF |
| | 16,80 | T12ER0 | |
| | 16,81 | A1EM43,74 | LETTRE A |
| | 16,82 | A2EM43,75 | LETTRE B. |
| | 16,83 | T1R9 | PREPARATION |
| | 16,84 | T2R10 | |
| | 16,85 | RCV26,0 | APPEL |
| | 16,86 | T1 0R3 | RANGEMENT RESULTAT |
| | 16,87 | T3M43,77 | LETTRE D. |
| | 16,88 | RR0 | |
| | 16,89 | X | |
| | 16,90 | X | |
| | 16,91 | X | |
| | 16,92 | X | |
| | 16,93 | X | |
| | 16,94 | X | |
| | 16,95 | X | |

# C H A P I T R E  II

## LA METHODE DE LAGUERRE

*Pour la résolution des équations $f(x) = 0$, il existe de nombreuses méthodes itératives. Nous rappelerons la méthode de Newton à convergence quadratique et étudierons une méthode analogue due à Laguerre, valable pour les seuls polynômes réels à solutions toutes réelles, mais possédant dans ce cas d'intéressantes propriétés. Elle donne lieu à une convergence monotone, cubique vers toute solution réelle simple. La convergence ne dépend pas de la valeur initiale choisie. Il est alors possible de définir deux suites qui convergent vers les solutions encadrant la valeur initiale (notion de zone d'exclusion), en utilisant, si cela est nécessaire, la notion de "bouclage à l'infini."*

### 2.I. - La méthode de Newton

La méthode de Newton, applicable à un polynôme P(x) ayant des racines réelles ou complexes, semble particulièrement bien adaptée à la recherche des valeurs propres d'une matrice tridiagonale car elle ne nécessite que le calcul des valeurs de P(x) et de sa première dérivée P' (x).

Les itérées successives de la méthode de Newton sont définies par :

$$(I) \qquad \boxed{x_{i+1} = x_i - \frac{P(x_i)}{P'(x_i)}} \qquad P'(x_i) \neq 0$$

La convergence est quadratique, et ne peut être assurée que dans le voisinage d'une solution simple, ce qui rend délicat l'emploi des techniques correspondantes. Cette difficulté a conduit Laguerre "*à trouver une méthode qui conduise de la façon la plus sûre et la plus rapide aux valeurs approchées des racines d'un polynôme donné*". (Oeuvres de Laguerre, Tome I, page 87 [9]).

Au lieu d'exposer cette méthode, comme le fait J.H. WILKINSON [ 10 ], il nous a paru souhaitable de la présenter de deux manières différentes en introduisant ou non le degré du polynôme et en détaillant dans chaque cas les propriétés.

## 2.2. - La méthode $\mathcal{L}_\infty$ de Laguerre

(Solutions réelles distinctes).

Soit P(x) un polynôme normé (c'est-à-dire dont le coefficient du terme de plus haut de ré est égal à 1) de degré n, ayant des solutions $r_k$ *que nous supposerons, provisoirement réelles, distinctes et classées dans l'ordre croissant* : $r_1 < r_2 \ldots < r_n$

$$(I) \quad P(x) = \sum_{k=1}^{n} (x-r_k)$$

$$(2) \quad \frac{P'(x)}{P(x)} = \sum_{k=1}^{n} \frac{1}{(x-r_k)}$$

$$(3)\quad \frac{P'(x)}{P(x)} \sim \frac{1}{x-r_j} \qquad \text{pour } x \sim r_j$$

$$(4)\quad r_j \sim x - \frac{P(x)}{P'(x)}$$

ce qui conduit de suite à une première formule d'itération (celle de Newton) :

$$(5)\quad \boxed{x_{i+1} = x_i - \frac{P(x_i)}{P'(x_i)} = x_i - 1/p \quad \text{pour } x_o \sim r_j}$$

où $p(x_i)$, noté $p$ pour alléger les notations lorsqu'il ne peut y avoir confusion, est défini par

$$(6)\quad p = \frac{P'(x_i)}{P(x_i)} = \sum_{k=1}^{n} \frac{1}{(x_i - r_k)}$$

PLUTOT QUE DE POURSUIVRE LE DEVELOPPEMENT AINSI AMORCE, LAGUERRE A PREFERE DERIVER LES DEUX MEMBRES DE LA FORMULE (2) ET ECRIRE :

$$(7)\quad \frac{P'^2(x) - P(x)\,P''(x)}{p^2(x)} = \sum_{k=1}^{n} \frac{1}{(x-r_k)^2} = q^2$$

q ainsi défini, sera toujours pris positif ou nul.

Deux inégalités fondamentales s'en déduisent de suite :

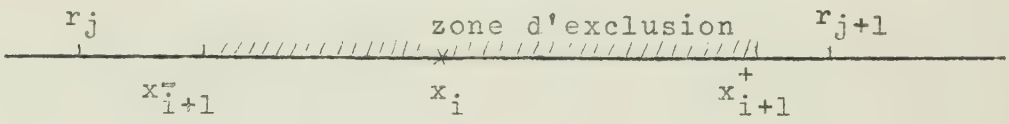
$$(8)\quad P'^2(x) - P(x)\,P''(x) \geq 0 \qquad \forall\, x$$

$$(9)\quad |x-r_k| \geq \frac{P}{\sqrt{P'^2 - PP''}} \qquad \forall\, x,k$$

*On a ainsi défini une zone d'exclusion des solutions (au sens de Householder) et un second procédé itératif que nous noterons $\mathcal{L}_\infty(x_i)$.*

$$(10)\qquad \boxed{x_{i+1} = x_i + \upsilon \frac{P(x_i)}{\sqrt{P'^2(x_i) - P(x_i)P''(x_i)}} = \mathcal{L}_{\infty}(x_i);\ \upsilon^2 = 1}$$

suivant le choix du signe de $\upsilon$, ce procédé fournit deux approximations de $r_j$ et de $r_{j+1}$ meilleures que $x_i$ lui-même, que nous noterons respectivement $x^-_{i+1}$ ($\upsilon P < 0$) et $x^+_{i+1}$ ($\upsilon P > 0$ c'est-à-dire signe ($\upsilon$) = signe (P)).




Plus généralement, en choisissant un signe constant de $\upsilon$, on définit deux suites $x^-_i$ ($\upsilon = -1$) et $x^+_i$ ($\upsilon = +1$) qui convergent de façon monotone vers les deux solutions $r_j$ et $r_{j+1}$ les plus voisines de la valeur initiale $x_o$. Lorsque $x_o$ est inclus dans l'intervalle contenant toutes les solutions et différent d'une solution, montrons cette convergence et du même coup son caractère cubique pour une solution réelle simple.

$x_o \in (r_1, r_n)$ et $\forall k$ , $x_o \neq r_k \Rightarrow \exists j$ : $x_o \in (r_j, r_{j+1}) \Rightarrow$

$\left[x^-_i,\ x^+_i\right] \subsetneq \left[r_j, r_{j+1}\right]$ . La relation (9) montre que $r_j \lesssim x^-_{i+2} \lesssim x^-_{i+1} \lesssim x_i \cdots \lesssim r_{j+1}$

En utilisant la variable $q(x)$ définie en (7), la formule (10) s'écrit plus simplement :

$$(11)\qquad \boxed{\begin{aligned} x^-_{i+1} &= x_i - 1/q(x_i) \\ x^+_i{+}1 &= x_i + 1/q(x_i) \end{aligned}}$$

Soit $r_j$ une solution du polynôme et $x_o > r_j$. Posons :

$$\varepsilon = x_i - r_j \Longrightarrow \varepsilon > 0 \tag{12}$$

$$C_j^2 = \sum_{\substack{k=1 \\ k \neq j}}^{n} \frac{1}{(x_i - r_k)^2} \tag{13}$$

Au voisinage de la solution $r_j$, nous avons :

$$q^2 = \frac{1}{\varepsilon^2} + C_j^2 \Longrightarrow \frac{1}{q} = \varepsilon - \varepsilon^3 \frac{C_j^2}{2} > 0$$

$$x_{i+1}^- - r_j = x_i - r_j - \frac{1}{q} \sim \varepsilon^3 \frac{C_j^2}{2}$$

$$\boxed{\frac{x_{i+1}^- - r_j}{(x_i - r_j)^3} \sim \frac{C_j^2}{2}} \tag{14}$$

. Lorsque la valeur initiale $x_o$ est à l'extérieur de l'intervalle $\left[r_1, r_n\right]$, l'une des deux suites diverge. Si $x_o > r_n$ la suite $x_i^-$ converge cubiquement vers $r_n$ mais la suite $x_i^+$, étant croissante, augmente indéfiniment. De plus, pour x suffisamment grand :

$$x_{i+1}^+ \sim \frac{\sqrt{n} + 1}{\sqrt{n}} \; x_i \tag{15}$$

et cette divergence assez lente nécessite l'utilisation d'un test d'arrêt (par exemple, une comparaison avec une bonne supérieure connue des solutions)

Si $x_o < r_1$, il en est de même en changeant les rôles de $x_i^+$ et $x_i^-$. Cet inconvénient n'existe pas dans le cas de la formule $\mathcal{L}_n$ que nous allons définir. Cette formule donnera lieu à une convergence quelque soit la valeur de $x_o$.

## 2.3. - La méthode $\mathcal{L}_n$ de Laguerre

(Solutions réelles distinctes)

On peut affiner le procédé utilisé dans la méthode $\mathcal{L}_\infty$ en définissant un intervalle d'exclusion plus grand que le précédent, à savoir :

(1) $\left[ x - 1/u_2,\ x - 1/u_1 \right)$

$u_1$ et $u_2$ étant les solutions ($u_1 < 0 < u_2$) de l'équation

(2) $\boxed{f(u) = n u^2 - 2 p u + p^2 - (n-1) q^2 = 0}$

n est le degré du polynôme, p et q sont les fonctions déjà définies :

(3) $$p = \sum_{k=1}^{n} \frac{1}{x-r_k} = \frac{P'(x)}{P(x)}$$

(4) $$q^2 = \sum_{k=1}^{n} \frac{1}{(x-r_k)^2} = \frac{P'^2(x) - P(x)\,P''(x)}{P^2(x)} \qquad (q > 0)$$

Nous allons démontrer successivement :

2.3.1. *qu'il s'agit d'un domaine d'exclusion.*

Si l'on convient de poser

(5) $$A_k = \frac{1}{x-r_k}$$

$$p = \sum_{k=1}^{n} \frac{1}{x-r_k} = \sum_{k=1}^{n} A_k$$

(6) $$p^2 = (\Sigma A_k)^2 \leq n \Sigma A_k^2 = nq^2$$ (inégalité de Cauchy Schwartz). En utilisant les formules (3) et (4) qui définissent p et q, on en déduit :

(7) $p'^2 \leq n (p'^2 - PP'')$.

Le Hessien $H(x) = (n-1)P'^2 - n\,P\,P''$ est donc toujours non négatif ainsi qu'il en résulte du reste de la théorie *des émanants, théorie qui domine la pensée de Laguerre*. En appliquant à nouveau l'inégalité de Cauchy Schwartz, on obtient :

$$(8) \qquad (p - A_k)^2 \leq (n-1)\,(q^2 - A_k^2).$$

Cette inégalité s'écrit également :

$$(9) \qquad f(A_k) = n\,A_k^2 - 2p\,A_k + (p^2 - (n-1)\,q^2) \leq 0$$

Le polynôme f(u) possède deux solutions réelles (son discriminant $\Delta = (n-1)\,(nq^2-p^2)$ est non négatif) que nous notons $u_1$ et $u_2$ :

$$(10) \qquad \boxed{\begin{aligned} u_1 &= \frac{1}{n}\left(p - \sqrt{(n-1)(nq^2-p^2)}\right) \\ u_2 &= \frac{1}{n}\left(p + \sqrt{(n-1)(nq^2-p^2)}\right) \end{aligned}}$$

Entre les solutions $u_1$ et $u_2$, f(u) est négatif d'où :

$$(11) \qquad \forall k : f(A_k) \leq 0 \quad ; \quad u_1 \leq A_k = (x-r_k)^{-1} \leq u_2.$$

Cette double inégalité étant vraie pour une valeur quelconque de k :

$x \in (r_1, r_n)$ $\quad x-r_1 > 0$, $\quad x-r_n < 0$ $\quad u_1 < 0 < u_2$ $\quad p^2 < (n-1)q^2$,

car les deux solutions de f(u) sont alors de signes contraires.

$$(12) \qquad \boxed{x \in (r_1, r_n) \Longrightarrow p^2 < (n-1)\,q^2}$$

Lorsque x est intérieur à l'intervalle $(r_1, r_n)$ c'est-à-dire lorsqu'il y a $\alpha$ racines inférieures à x et $\beta$ supérieures ($\alpha > 0$, $\beta > 0$), on peut affirmer plus généralement :

$$(13)\quad \boxed{x \in (r_1, r_n) \Longrightarrow p^2 < \max(n-\alpha,\ n-\beta)q^2}$$

Nous nous réservons d'étudier plus loin le cas où x est extérieur à l'intervalle contenant les solutions.

En appelant $r_j$ et $r_{j+1}$ les solutions de P(x) les plus proches de x et l'encadrant ($r_j \leq x \leq r_{j+1}$), on déduit de (11) :

$$x - r_j \geq \frac{1}{u_2} \qquad x - r_{j+1} \leq \frac{1}{u_1}$$

$$(14)\quad r_j \leq x - \frac{1}{u_2} < x < x - \frac{1}{u_1} \leq r_{j+1}$$

La considération de ce domaine d'exclusion permet de définir la seconde formule d'itération de Laguerre que nous notons $\mathcal{L}_n(x_i)$ $\qquad u^2 = 1$

$$(15)\quad \boxed{x_{i+1} = \mathcal{L}_n(x_i) = x_i - \frac{n\,P(x_i)}{P'(x_i) + u\sqrt{(n-1)\,[(n-1)P'^2(x_i) - nP(x_i)P''(x_i)]}}}$$

2.3.2 - *Que ce domaine d'exclusion est plus grand que celui défini pour* $\mathcal{L}_\infty$

$$(16)\quad \left[x - \frac{1}{q},\ x + \frac{1}{q}\right] \subset \left[x - \frac{1}{u_2},\ x - \frac{1}{u_1}\right],$$

autrement dit :

$$(17)\quad \left|\frac{1}{u_1} - \frac{1}{u_2}\right| = \frac{2\sqrt{(n-1)(nq^2-p^2)}}{(n-1)q^2 - p^2} > \frac{2}{q}$$

$$q\sqrt{(n-1)(nq^2-p^2)} > (n-1)q^2 - p^2$$

$$(18)\quad (n-1)q^4 + (n-1)p^2q^2 - p^4 > 0.$$

Considérons le premier membre de (18) comme un polynôme du second degré en $p^2$ que nous notons $P(p^2)$. Les solutions en sont réelles et de signes opposés (son discriminant $\Delta = (n-1)(n+3)q^4$ est non négatif, le produit des solutions : $-(n-1)q^4$ est non positif). $P(x)$ est donc positif entre les solutions. On déduit de (12) : $x \in (r_1, r_n) \Longrightarrow p^2 \leq (n-1)q^2 \Longrightarrow p^2 \leq \frac{1}{2}\left[\sqrt{(n-1)(n+3)}+(n-1)\right] q^2$; $p^2$ est donc inférieur à la solution positive $(2n-2 < \sqrt{(n-1)(n+3)}+n-1)$ et les relations (18) et (16) sont alors vérifiées.

2.3.3. *Que ce domaine tend vers le domaine défini pour $\mathcal{L}_\infty$ lorsque $n \to \infty$*

En effet :

$$(19) \quad \frac{2\sqrt{(n-1)(nq^2-p^2)}}{(n-1)q^2 - p^2} \longrightarrow \frac{2}{q} \text{ quand } n \to \infty$$

Ceci justifie les notations que nous employons. A la première exclusion, nous faisons correspondre $x_{i+1} = \mathcal{L}_\infty(x_i)$, et à la seconde, $x_{i+1} = \mathcal{L}_n(x_i)$, où n est le degré du polynôme $P(x)$. *Il est possible de définir plus généralement une formule d'itération $\mathcal{L}_N$, obtenue en remplaçant n par $N \geq n$. Cette formule possède les mêmes propriétés que $\mathcal{L}_n$ mais "marche moins bien si le degré maximal N est plus loin du vrai degré n du polynôme à étudier".* (conférence de A.M. OSTROWSKI - Besançon - 22 mai 1964). Aussi nous bornerons-nous à étudier $\mathcal{L}_n$. Pour que les notations soient concordantes, on posera alors :

(20)
$$\boxed{\begin{aligned} x_{i+1}^{-} &= \mathcal{L}_n(x_i) = x_i - \frac{1}{u_2} \text{ avec } u_2 = \frac{1}{n}\left(p + \sqrt{(n-1)(nq^2-p^2)}\right) \\ x_{i+1}^{+} &= \mathcal{L}_n(x_i) = x_i - \frac{1}{u_1} \text{ avec } u_1 = \frac{1}{n}\left(p - \sqrt{(n-1)(nq^2-p^2)}\right) \end{aligned}}$$

2.3.4. *La formule d'itération $\mathcal{L}_n(x)$ donne aussi une convergence cubique*

Démontrons par exemple que $x_{i+1}^{-} = \mathcal{L}_n(x_i) = x_i - 1/u_2$ converge cubiquement vers la solution inférieure la plus proche $r_j$.

$$r_1 \leq r_j \leq x_{i+1}^{-} \leq x_i < x_{i+1}^{+} \leq r_{j+1} \leq r_r$$

convenons de poser

(21) $\quad \varepsilon = x_i - r_j \Longrightarrow \varepsilon > 0$

(22) $\quad B_j = \sum_{\substack{k=1 \\ k\neq j}}^{n} \frac{1}{(x-r_k)} = \sum_{\substack{k=1 \\ k\neq j}}^{n} \Lambda_k$

(23) $\quad C_j^2 = \sum_{\substack{k=1 \\ k\neq j}}^{n} \frac{1}{(x-r_k)^2} = \sum_{\substack{k=1 \\ k\neq j}}^{n} \Lambda_k^2$

Comme nous l'avons déjà fait pour les formules correspondantes (2.2.12) et(2.2.13). Alors :

$$p = \frac{1}{\varepsilon} + B_j \qquad q^2 = \frac{1}{\varepsilon^2} + C_j^2$$

$$u_2 = \frac{1}{n}\left(\frac{1}{\varepsilon} + B_j + \sqrt{(n-1)\,\frac{n-1}{\varepsilon^2} - \frac{2B_j}{\varepsilon} + nC_j^2 - B_j^2}\,\right)$$

$$= \frac{1}{n\varepsilon}\left(1 + \varepsilon B_j + (n-1)\sqrt{1 - \frac{2B_j\varepsilon}{n-1} + \frac{nC_j^2 - B_j^2}{n-1}\,\varepsilon^2}\,\right)$$

$$\sim \frac{1}{n\varepsilon}\left(n + \frac{n}{2(n-1)}\left((n-1)C_j^2 - B_j^2\right)\varepsilon^2\right)$$

$$x_{i+1}^{-} - r_j \sim \varepsilon^3\,\frac{(n-1)C_j^2 - B_j^2}{2(n-1)}$$

$$(24)\qquad \boxed{\frac{x_{i+1}^{-} - r_j}{(x_i - r_j)^3} \sim \frac{C_j^2}{2} - \frac{B_j^2}{2(n-1)}}$$

Lorsque $x \in (r_1, r_n)$, on vérifie immédiatement que $B_j^2 \leq (n-1)\, C_j^2$. On a donc une convergence cubique plus rapide que pour $\mathcal{L}_\infty$.

Lorsque la valeur initiale $x_o$ est extérieure à l'intervalle $\left[r_1, r_n\right]$ contenant toutes les solutions de $P(x)$, il y a toujours convergence, mais il nous faut distinguer, comme nous l'avons déjà fait pour $\mathcal{L}_\infty$, les suites $x_i^-$ et $x_i^+$ et préciser leur convergence.

Remarquons au préalable que, au voisinage de l'infini :

$$p^2-(n-1)q^2 = P^2\left[(n-1)PP'' -(n-2)P'^2\right]$$

$$\sim P^2\left[(n-1)n(n-1)x^{2n-2} - (n-2)n^2x^{2n-2}\right] = nP^2x^{2n-2} > 0$$

D'autre part (formule 2.3.12)

$$x \in \left[r_1, r_n\right] \Longrightarrow p^2 - (n-1)q^2 < 0$$

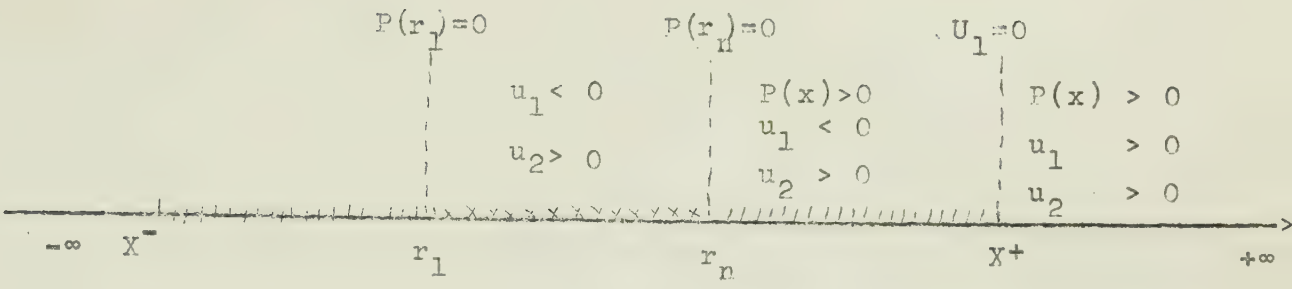
On en déduit que les solutions de $p^2-(n-1)q^2 = 0$ existent et sont extérieures à l'intervalle $\left[r_1, r_n\right]$.

Soit $X^+$ une solution de $p^2-(n-1)q^2=0$ telle que $X^+ > r_n$ et précisons la convergence pour $x_o > r_n$. Les solutions $u_1$ et $u_2$ du polynôme $f(u)$ défini en (2.3.2.) vérifient :

$$x \in (r_n, X^+) \Longrightarrow u_1(x) < 0 < u_2(x)$$

$$u_1(X^+) = 0 \qquad u_2(X^+) = 2n/n > 0$$

$$x > X^+ \Longrightarrow 0 < u_1(x) < u_2(x)$$




Nous allons montrer successivement :

LEMME (a) : $r_n < x_i \leq X^+ \Longrightarrow$ suite $x_i^+$ croissante

en effet :

$$x_{i+1}^+ = x_i^+ - 1/u_1 > x_i^+ \text{ puisque } u_1(x_i^+) < 0$$

LEMME (b) : $r_n < x_i \leq X^+ \Longrightarrow \dfrac{x_{i+1}^+}{x_i^+} > 1+\alpha$ ; $\left(\alpha = \dfrac{(x_o - r_n)^2}{\sqrt{n-1}\, X^+} > 0\right)$

$$\frac{x_{i+1}^+}{x_i^+} = 1 + \frac{1}{|x_i|\, x_i^+} > 1 + \alpha \Longleftrightarrow |u_1|\, x_i^+ < \beta = 1/\alpha$$

Or :

$$n|u_1| = -p + \sqrt{(n-1)(nq^2 - p^2)} < \sqrt{(n-1)(nq^2 - p^2)}$$

$$< q\sqrt{n-1} = \sqrt{n-1} \sum_k (x_i - r_k)^{-2} \leq n\sqrt{n-1}\,(x_i - r_n)^{-2} \leq n\sqrt{n-1}\,(x_o - r_n)^{-2}$$

d'où $\beta = \sqrt{n-1}\, X^+ (x_o - r_n)^{-2}$

LEMME (c) : $\exists \tau$ : $x_\tau > X^+ \Longrightarrow x_{\tau+1}^+ < r_1$.

En effet : $x_\tau > X^+ \Longrightarrow p(x_\tau)^2 - (n-1)q^2(x_\tau) > 0$

$\Longrightarrow 0 < u_1(x_\tau) < (x_\tau - r_1)^{-1} < (x_\tau - r_n)^{-1} < u_2(x_\tau)$

puisque (formule 2.3.9) $\Lambda_1 = (x_\tau - r_1)^{-1}$ et $\Lambda_n = (x_\tau - r_n)^{-1}$ sont comprises entre les solutions positives $u_1(x_\tau)$ et $u_2(x_\tau)$ du polynôme $f(u)$ défini pour $x_\tau$.

On en déduit immédiatement :

$$x_\tau > X^+ \Longrightarrow \begin{cases} x_\tau - r_1 < u_1^{-1} \Longrightarrow x^+_{\tau+1} = x_\tau - u_1^{-1} < r_1 \\ x_\tau - r_n > u_2^{-1} \Longrightarrow x^-_{\tau+1} = x_\tau - u_2^{-1} > r_n \end{cases}$$

LEMME (d) : Pour $x_o > r_n$, les deux suites $x_i^+$ (au dela des $\tau$ premières itérations) et $x_i^-$ convergent cubiquement vers les solutions $r_1$ et $r_n$. On utilise pour cela une démonstration analogue à celle présentée dans le cas $x_o \in (r_1, r_n)$. Il est possible de montrer de la même manière que $x_o < r_1 \Longrightarrow \begin{cases} x_i^- \to r_n & \text{(cubiquement au delà des } \tau' \text{ premières itérations)} \\ x_i^+ \to r_1 & \text{(cubiquement).} \end{cases}$

Ces résultats sont représentés figure (3).

## 2.4. - Cas des Solutions réelles multiples

Nous supposons maintenant que les solutions $r_k$ du polynôme $P(x)$ sont réelles et rangées en ordre croissant :

$$r_1 \leq r_2 \leq \ldots \leq r_{n-1} \leq r_n$$

Soit $r_j$ une solution multiple et $m$ sa multiplicité, telle que :

$$r_j = r_{j+1} = \ldots = r_{j+m}$$

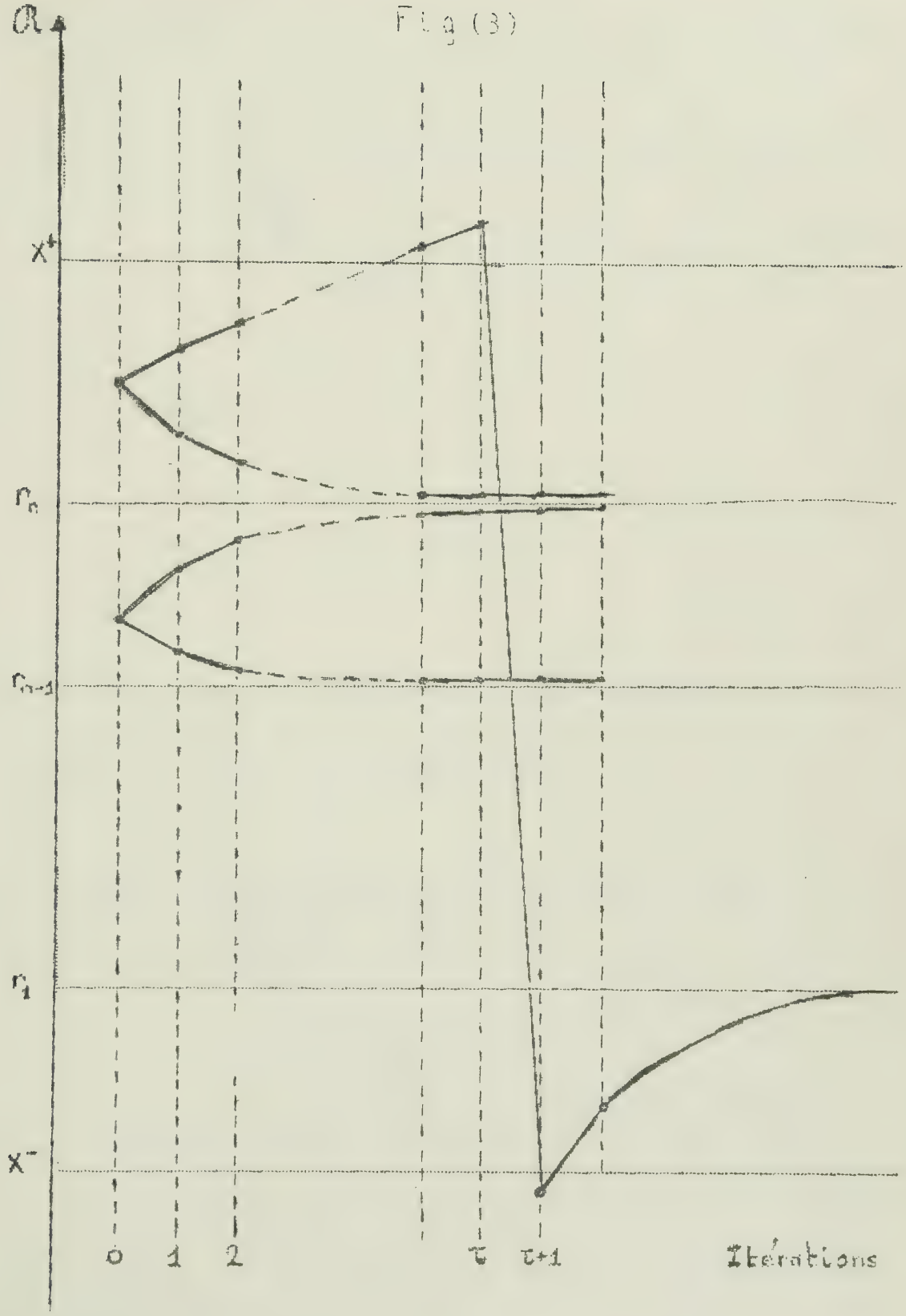
Etudions dans ce cas, le type de convergence obtenu pour les formules de Laguerre précédemment définies. Pour cela posons :

$$(1) \quad \varepsilon = x - r_j \qquad (x > r_j \Longrightarrow \varepsilon > 0)$$

$$(2) \quad B_j = \sum_{\substack{k=1 \\ k \neq j \ldots k \neq j+m}}^{n} \frac{1}{(x - r_k)}$$

$$(3) \quad C_j^2 = \sum_{\substack{k=1 \\ k \neq j \ldots k \neq j+m}}^{n} \frac{1}{(x - r_k)^2}$$

Fig (3)



généralisant ainsi les variables $B_j$ et $C_j^2$ définies ci-dessus pour les solutions distinctes.

2.4.1. Cas de $\mathcal{L}_\infty$

Au voisinage de la solution multiple $r_j$ $(x > r_j)$ nous avons :

$$(4)\quad q^2 = \sum_{k=1}^{n} \frac{1}{(x - r_k)^2} = \frac{m}{\varepsilon^2} + C_j^2$$

$$\frac{1}{9} \sim \frac{\varepsilon}{\sqrt{m}} \quad (1 - \varepsilon^2 \quad \frac{C_j^2}{2m}$$

$$\mathcal{L}_\infty(x) \sim \varepsilon \quad (1 - \frac{1}{\sqrt{m}})$$

$$(5)\quad \frac{x_{i+1} - r_j}{x_i - r_j} \sim 1 - \frac{1}{\sqrt{m}}$$

On ne peut affirmer plus que la convergence linéaire.

Il en va de même lorsque x est proche d'une solution multiple qui lui est supérieure.

2.4.2 Cas de $\mathcal{L}_n$

Dans les mêmes conditions que pour $\mathcal{L}_\infty$, nous avons

$$(6)\quad p = \frac{m}{\varepsilon} + B_j$$

$$(7)\quad q^2 = \frac{m}{\varepsilon^2} + C_j^2$$

$$u_2 = \frac{1}{n} \; p + \sqrt{(n-1)\,(nq^2 - p^2)}$$

$$u_2 \sim \frac{1}{n} \; (m + B_j \varepsilon + \sqrt{(n-1)\,\left[m(n-m) - 2m\, B_j \varepsilon\right]}$$

$$u_2 \sim \frac{1}{n\varepsilon} \left[ m + \sqrt{m(n-1)\ (n-m)} + B_j \left(1 - \sqrt{\frac{m\ (n-1\ )}{n-m}}\ \varepsilon\right)\right] \quad II_{(I5)}$$

$$\frac{1}{u_2} \sim \frac{n}{m + \sqrt{m\ (n-1\ )\ (n-m)}}\ \varepsilon$$

$$(8) \qquad \frac{x_{i+1} - r_j}{x_i - r_j} \sim 1 - \frac{n}{m + \sqrt{m\ (n-1)\ (n-m)}}$$

On ne peut affirmer plus que la convergence linéaire. Il en va encore de même lorsque $x_o$ est inférieur à une solution multiple voisine.

Comme pour la convergence cubique, nous vérifions que la convergence obtenue par $\mathcal{L}_n$ sera plus rapide que celle correspondant à $\mathcal{L}_\infty$ ; en effet :

$$1 - \frac{1}{\sqrt{m}} > 1 - \frac{n}{m + \sqrt{m\ (n-1\ )\ (n-m)}}$$

$$n\sqrt{m} > m + \sqrt{m\ (n-1\ )\ (n-m)}$$

$$m + 1 > 2\sqrt{m}$$

Ostrowski a étudié la possibilité d'améliorer les formules de Laguerre dans le cas des solutions multiples mais cela présente peu d'intérêt pratique car cette amélioration suppose connue l'ordre de multiplicité, ce qui n'est pas en général réalisé.

## 2.5. Calcul numérique d'une solution d'un polynôme à solutions réelles selon la méthode de Laguerre.

*Une particularité importante de la méthode de Laguerre est d'obtenir, par le simple choix d'un signe, une convergence monotone sur la droite réelle munie d'un seul point à l'infini (bouclage à l'infini, par exemple* $x_n^+ > x^+ \neq x_{n+1}^+ < r_1$ *). Plus exactement les deux suites* $x_i^+$ *et* $x_i^-$ *tendent de manière monotone vers les deux solutions* $r_j$ *et* $r_{j+1}$ *encadrant (au sens de la droite déjà définie)* $x_o$ ($x_o \neq r_k \ \forall k$).

### 2.5.1 Règle des Signes

On obtient la convergence vers la solution supérieure la plus proche en choisissant pour $\upsilon$ le même signe que P ( $\upsilon P > 0$) et vers la solution inférieure la plus proche en choisissant pour $\upsilon$ le signe de - P ( $\upsilon P < 0$).

Démonstration pour $\mathcal{L}_n$

1°/ Supposons $x_o \in [r_1, r_n]$

Nous avons vu que les corrections $\dfrac{n\,P(x_o)}{P'(x_o) - \sqrt{H(x_o)}}$ et $\dfrac{n\,P(x_o)}{P'(x_o) + \sqrt{H(x_o)}}$ proposées par Laguerre étaient de signes opposés. On en déduit : $(P' - \sqrt{H})^{-1} < 0 < (P' + \sqrt{H})^{-1}$

Comme la formule $\mathcal{L}_n$ s'écrit : ($\upsilon^2 = 1$)

$$\mathcal{L}_n(x_o) = x_o - \frac{n\,P(x_o)}{P'(x_o) - \upsilon\sqrt{(n-1)^2\,P'^2(x_o) - n(n-1)P(x_o)\,P''(x_o)}}$$

On vérifie que le choix de $\upsilon = -1$ nous fait converger vers la solution $r_j$ inférieure si $P > 0$ et $r_{j+1}$ supérieure si $P < 0$ .

2°/ Supposons maintenant $x_o \notin \left[ r_1 , r_n \right]$ et par exemple $x_o > r_n$ . Les résultats obtenus s'étendront aisément au cas $x_o < r_1$.

Le polynôme étant supposé normé ($a_o = 1$ ) nous avons toujours $P(x_o) > 0$

Pour $x_o < X^+$ , les corrections sont de signesopposés (comme dans le cas ci-dessus) :

$\upsilon = 1 \Rightarrow x_i^+$ croît tant que $x_i < X^+$

$\upsilon = -1 \rightarrow x_i^- \rightarrow r_n$

Lorsque $x_o > X^+$ , les corrections sont alors de même signe, et

$$0 < \frac{n\,P(x_o)}{P'(x_o) + \sqrt{H(x_o)}} < \frac{n\,P(x_o)}{P'(x_o) - \sqrt{H(x_o)}}$$

La correction la plus petite, obtenue en choisissant $\upsilon = -1$, permet de converger vers $r_n$. Par contre le choix de $\upsilon = 1$ entraîne un bouclage par l'infini.

$\upsilon = -1 \quad \Rightarrow \quad x_i^- \rightarrow r_n$

$\upsilon = 1 \quad \Rightarrow \quad x_1^+ < r_1 \quad \Rightarrow \quad x_i^+ \rightarrow r_1$ .

Démonstration pour $\mathcal{L}_\infty$

$$\mathcal{L}_\infty(x_o) = x_o + \upsilon \frac{P(x_o)}{\sqrt{P'(x_o)^2 - P(x_o)\,P''(x_o)}}$$

Lorsque $x_o$ est à l'intérieur de l'intervalle contenant toutes les solutions, les termes de corrections obtenus pour $\upsilon = 1$ ou $\upsilon = -1$ sont de signes opposés.

Si $\upsilon P > 0$, la suite $x_i^+$ converge vers la solution supérieure la plus proche. Si $\upsilon P < 0$, la suite $x_i^-$ converge vers la solution inférieure la plus proche. Lorsque $x_o$ est à l'extérieur de l'intervalle $\left[r_1 , r_n\right]$, nous avons vu que l'une des suites diverge ; on peut cependant préciser que

$$x_o > r_n \quad , \quad \upsilon P < 0 \quad \rightarrow \quad x_i^- \rightarrow r_n$$

$$x_o < r_1 \quad , \quad \upsilon P > 0 \quad \rightarrow \quad x_i^+ \rightarrow r_1$$

<u>En résumé</u> - La suite $x_i^+$ est obtenue en choisissant $\upsilon P > 0$. Elle donne lieu à une convergence vers la droite (en direction de $+\infty$) ($x_i^+ < x_{i+1}^+$) sauf pour le bouclage à l'infini déjà signalé) Il en est de même de la suite $x_i^-$, obtenue en choisissant $\upsilon P < 0$ et qui donne lieu, dans les mêmes conditions, à une convergence dans la direction de $-\infty$ ($x_{i+1}^- < x_i^-$). Cette remarque justifie les notations utilisées pour ces suites.

2.5.2 <u>Choix d'une valeur initiale</u>

La convergence est indépendante du choix de la valeur initiale $x_o$ mais les deux suites $x_i^+$ et $x_i^-$, définies par $x_o$, convergent vers les deux solutions encadrant $x_o$.

Si l'on désire, *calculer la solution de plus petit module*, on prendra comme valeur initiale $x_o = 0$ et on calculera les solutions, à droite et à gauche de $x_o = 0$.

*Pour calculer la solution de plus grand module*, ou, plus généralement $r_1$ et $r_n$, on peut utiliser comme valeur initiale, une des bornes de l'intervalle $\left[ r_1 , r_n \right]$, obtenue en étudiant la méthode $\mathcal{L}_n$ au voisinage de l'infini. On sait en effet que $X^+$ et $X^-$, ont une valeur finie.

Les deux valeurs correpondant à la formule $\mathcal{L}_n$ pour $x \sim \infty$, sont alors deux bornes encadrant l'intervalle $\left[ r_1 , r_n \right]$ et que nous allons préciser :

Soit P(x) le polynôme de degré n :

$$P(x) = x^n + a_1 x^{n-1} + a_2 x^{n-2} + \cdots\cdots + a_n$$

$$P'(x) = n x^{n-1} + (n-1) a_1 x^{n-2} + (n-2) a_2 x^{n-3} + \cdots + a_{n-1}$$

$$P''(x) = n(n-1)x^{n-2} + (n-1)(n-2) a_1 x^{n-3} + (n-2)(n-3) a_2 x^{n-4} + \cdots\cdots + a_{n-2}$$

Au voisinage de l'infini, le Hessien :

$$H(x) = (n-1)^2 P'(x)^2 - n (n-1) P(x) P''(x)$$

est équivalent à :

$$H(x) \sim (n-1) \left[ (n-1) a_1^2 - 2 n a_2 \right] x^{2n-4}$$

Lorsque P(x) ne possède que des solutions réelles, le terme $(n-1) \left[ (n-1) a_1^2 - 2 n a_2 \right] x^{2n-4}$ équivalent au Hessien est toujours positif.

En effet :

$$a_1 = -\sum_{k=1}^{n} r_k$$

$$a_2 = \sum_{j=1}^{n}\sum_{k=j}^{n} r_j r_k = \frac{1}{2}\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n} r_j r_k$$

$$= \frac{1}{2}\left[(\sum_{k=1}^{n} r_k)^2 - \sum_{k=1}^{n} r_k^2\right]$$

$$(\sum_{k=1}^{n} r_k)^2 \leq n \sum_{k=1}^{n} r_k^2 \quad \text{inégalité de Cauchy Schwartz}$$

d'où :

$$(n-1)a_1^2 = (n-1)(\sum_{k=1}^{n} r_k)^2 \geq n \sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n} r_j r_k = 2\,n\,a_2$$

La formule $\alpha_n(x) = x - \dfrac{nP(x)}{P'(x) - \upsilon\sqrt{H(x)}}$ calculée pour x voisin de l'infini, donne deux valeurs suivant le choix de $\upsilon$ ($\upsilon = 1$ ou $\upsilon = -1$) qui sont les bornes cherchées.

$$b^- = \frac{1}{n}\left(-a_1 - \sqrt{(n-1)\left[(n-1)a_1^2 - 2\,n\,a_2\right]}\right)$$

$$b^+ = \frac{1}{n}\left(-a_1 + \sqrt{(n-1)\left[(n-1)a_1^2 - 2\,n\,a_2\right]}\right)$$

$$x \sim \infty \Longrightarrow b^- < r_1 < r_n < b^+$$

Remarque I

La formule $\alpha_\infty(x)$ ne possède pas cette propriété intéressante :

$$x \sim \infty \Longrightarrow \alpha_\infty(x) \sim \left(1 + \frac{\upsilon}{\sqrt{n}}\right)x \quad \text{(Formule 2.3.I5)}$$

$$\forall\,\upsilon \quad (\upsilon^2 = 1) \quad x \to \infty \Longrightarrow \left(1 + \frac{\upsilon}{\sqrt{n}}\right)x \to \infty$$

Remarque 2 - Ayant déterminé le polynôme caractéristique d'une matrice M, il est facile de calculer les solutions les plus grandes et plus petites en module, quantités qui interviennent dans la définition du conditionnement de M.

2.5.3. Tests d'Arrêt

Comme pour toutes méthodes itératives, il est nécessaire d'arrêter les itérations lorsque la précision est jugée suffisante. Pour cela on vérifiera que l'écart (défini au chapitre I) entre deux itérées successives, e ($x_i$, $x_{i+1}$) est inférieur à un nombre donné indiquant la précision demandée. Par suite des erreurs de calcul, il se peut que ce test ne soit jamais vérifié et que l'on saute involontairement une solution. On peut alors prévoir aussi un second test vérifiant que, P étant très petit, le produit P P' change de signe. Ceci est facile à réaliser puisque l'on possède à tout moment les valeurs de P, P' et P".

D'autre part, nous avons déjà signalé (paragraphe 2.2) la nécessité d'un test lorsque l'on utilise la formule $\mathcal{L}_\infty$ car celle-ci ne possède pas le bouclage par l'infini. Pour éviter le calcul d'une suite divergente, on comparera chaque itérée aux bornes qui ont été définies ci-dessus.

En fait les tests à considérer dépendant aussi du balayage de l'axe réel que nous nous proposons de faire. Nous reviendrons donc sur cette question essentielle au paragraphe 2.6.2.

## 2.5.4. Coût d'une itération

Quelques soient les formules $\mathcal{L}_\infty$ ou $\mathcal{L}_n$ utilisées, il faut calculer P , P', P" (coût 3 Horners), et tenir compte du nombre d'opérations utilisées dans chacune des formules.

| Pour $\mathcal{L}_\infty$ | Pour $\mathcal{L}_n$ |
|---|---|
| 3 Horners | 3 Horners |
| I Racine carrée | I racine carré |
| 2 multiplications | 5 multiplications |
| I Division | I Division |
| 2 Additions | 3 Additions |

Il est donc légèrement plus rapide de calculer $\mathcal{L}_\infty$ , mais il est possible d'améliorer le calcul de $\mathcal{L}_n$, en utilisant les polynômes normés $\mathcal{P}$, $\mathcal{P}'$, $\mathcal{P}'$ associés à P, P', P".

$$\mathcal{P} = P / a_o = P \quad (a_o = 1)$$

$$\mathcal{P}' = P' / n$$

$$\mathcal{P}' = P'' / n(n-1)$$

Pour calculer les coefficients de $\mathcal{P}'$ et $\mathcal{P}'$ connaissant ceux de P, il faut (n-2) et (n-3) divisions supplémentaires.

En utilisant ces polynomes normés associés, les formules de Laguerre deviennent

$$\mathcal{L}_n(x) = x - \frac{\mathcal{P}}{\mathcal{P}' \pm \upsilon (n-1)\sqrt{\mathcal{P}'^2 - \mathcal{P}\mathcal{P}'}}$$ au lieu de

$$x - \frac{n P}{P' \pm \upsilon\sqrt{(n-1)^2 P'^2 - n(n-1) PP''}}$$

$$\alpha_\infty(x) = x + \frac{\upsilon P}{\sqrt{n^2 P^2 - n\,(n-1)\,P P'}} \qquad \text{au lieu de} \quad {}^{II}_{(23)}$$

$$x + \frac{\upsilon\ P}{\sqrt{P'^2 - P\ P''}}$$

La suppression de 2 multiplications dans le calcul de $\alpha_n$ est une amélioration intéressante si le nombre N d'itérations est supérieure à n-2 (on suppose pour cela que le temps d'une division est au moins égal à celui d'une multiplication.)

## 2.6. Algorithmes pratiques de calcul des Solutions d'un polynôme

Nous proposons, dans ce paragraphe, deux algorithmes pouvant être utilisées à la détermination de toutes les solutions d'un polynôme à solutions réelles. Le premier est une application à la méthode de Laguerre, d'un algorithme généralement utilisé avec les méthodes de Newton ou de dichotomie. Le second est valable essentiellement pour la méthode de Laguerre car il utilise la propriété de convergence monotone, propre à cette méthode.

Dans ces deux algorithmes, nous utiliserons, comme valeur initiale, l'une des itérées de l'infini précedemment définies (2.5. )

$$x_o = \frac{-a_1 - R}{n} \quad \text{avec} \quad R = \sqrt{(n-1)^2 a_1^2 - 2\,n(n-1)\,a_2}$$

et nous supposerons l'existence d'un balayage de l'axe réel dans le sens $(-\infty\ ,\ +\infty\ )$ c'est-àdire le calcul des suites $x_i^+$.

$$x_{i+1}^{+} = \alpha_n(x_i) = x_i - \frac{n\,P(x_i)}{P'(x_i) - \upsilon\sqrt{H(x_i)}}$$

avec $H(x_i) = (n-1)^2 P'(x_i)^2 - n(n-1) P(x_i)P''(x_i)$

Le choix du signe de $\upsilon$ est alors très simple :

$$\left.\begin{array}{l} \upsilon = -1 \quad \text{si } P < 0 \\ \upsilon = +1 \quad \text{si } P > 0 \end{array}\right\} \upsilon P > 0$$

2.6.I. Méthode par division successive

Cet algorithme est assez semblable à celui défini par B. Parlett [ 11 ], dans le cas de polynômes ayant des solutions toutes réelles.

Apliquant la formule de Laguerre à la valeur initiale et successivement aux valeurs $x_i^+$ obtenues, la suite $x_i^+$ converge vers la solution supérieure la plu proche

Au lieu de la série de tests proposée en toute généralité par B. Parlett, on peut, dans le cas qui nous intéresse, se contenter de la seule vérification suivante : L'écart relatif entre deux itérées successives est-il inférieur à la précision demandée ? (Théoriquement il suffirait de s'arrêter lorsque l'écart relatif est inférieur à la racine cubique de la précision relative en ordinateur, puisque la convergence est cubique.

L'expérience montre que ce test ne serait pas suffisant).

Ayant obtenu une solution $r^*_j$, nous divisons le polynôme par $(x - r^*_j)$ pour obtenir un polynôme de degré n- 1 possédant les mêmes solutions au même ordre de multiplicité sauf celle précédemment calculée dont l'ordre est diminué d'1 unité. Le schéma de Horner permet de calculer facilement les coefficients du polynôme de degré n- 1 . En effet :

Appelons $P_i$ les valeurs de Horner définies par :

$$P_i(x) = x\, P_{i-1}(x) + a_i$$

ou $P_o = a_o$ et $P_n = P(x)$

Pour une solution $r_j$, $P_n = P(r_j) = 0 \rightarrow xP_{n-1}(r_j) + a_n = 0$

Dans ces conditions, la division du polynôme par $(x - r_j)$ a comme reste $x\, P_{n-1}(r_j) + a_n = 0$. Les coefficients du polynôme-quotient sont alors les $P_i$.

Connaissant les coefficients du polynôme de degré n- 1 il est possible de déterminer une nouvelle valeur initiale et de reprendre le calcul jusqu'à ce que toutes les solutions aient été obtenues.

Un tel algorithme n'utilisant pas les propriétés propres à la méthode de Laguerre et, en particulier, le caractère monotone de la convergence, n'est pas très efficace et présente des inconvénients majeurs.

a) Une solution multiple sera calculée autant de fois que l'ordre de multiplicité avec une convergence linéaire. Il est difficile de supprimer cet inconvénient dans le cas général car un test sur l'ordre de multiplicité est délicat à définir.

b) A chaque détermination d'une solution, il faut calculer une valeur initiale et les coefficients du polynôme, augmentant le temps de calcul et créant une accumulation d'erreur préjudiciable à la précision, surtout sur les dernières solutions calculées.

c) Des dépassements de capacité peuvent être obtenus lors du calcul de P(x) pour une valeur trop grande de la variable x ( $P(x) > \infty^*$ ).
On peut alors utiliser une variante de cet algorithme où l'on choisit zéro comme valeur initiale et par laquelle on calculera les solutions dans un ordre approximativement croissant en module, en excluant le cas où zéro est une solution (alors P(0) = 0 ), les itérées de zéro sont $x_1^+$ et $x_1^-$ . On décidera de mener la convergence par la suite $x_i$ correspondante à la valeur $x_1^+$ ou $x_1^-$ la plus petite en module.

Cette variante permet de calculer des solutions très grandes en valeur absolue mais n'améliore pas la précision.

Nous proposons donc une méthode un peu différente, valable pour les polynômes à solutions réelles, plus précise et utilisant le caractère monotone de la convergence obtenue par la méthode de Laguerre.

2.6.2. Méthode par Saut

Deux propriétés essentielles de la Méthode de Laguerre (possibilité de choisir le sens de convergence et existence d'une convergence au moins linéaire, indépendante de la valeur initiale) suggèrent cependant d'adopter une méthode différente, la méthode par Saut, mieux adaptée à la recherche précise des solutions réelles d'un polynôme:

On s'impose un sens de balayage (et nous adopterons, pour unifier l'exposé, le sens - ∞ , + ∞ ) pour analyser la droite des réels entre deux bornes simples préalablement calculées et encadrant toutes les solutions réelles du polynôme P(x) étudié. Lorsqu'une solution réelle $r_j$ a été déterminée en utilisant la formule $\mathcal{L}_n$ de Laguerre, on continue le balayage dans le sens choisi en "sautant" la solution calculée. On évite ainsi un grand nombre de division de polynôme, source d'erreurs, mais cette méthode nécessite quelques remarques que nous allons préciser.

Remarque I - Choix des Bornes

On prendra comme bornes, les itérées de l'infini définies ci-dessus : l'une sera utilisée comme valeur initiale (borne inférieure), l'autre permettra de limiter le balayage (borne supérieure puisque nous avons choisi un balayage vers la droite).
Il est cependant impossible d'utiliser des bornes pour lesquelles le polynôme a une valeur supérieure à l'infini-machine (dépassement de capacité) et cette difficulté sera détectée différemment suivant l'ordinateur utilisé. Si ce dernier signale tout dépassement de capacité, on procèdera dès alerte donnée au calcul des solutions de module croissant en prenant zéro pour valeur initiale comme nous l'avons déjà indiqué (voir la méthode par division) Lorsque le dépassement de capacité n'est pas indiqué ce qui est le cas notamment pour la CAB 500, le nombre correspondant est automatiquement remplacé par l'infini-machine. CELA PEUT CONDUIRE A "OUBLIER" PLUSIEURS SOLUTIONS DE GRANDS MODULES. *Il est possible, d'autre part, que le bouclage par l'infini ait lieu pour une valeur de la variable inférieure à la borne supérieure. On prévoiera*

*donc un test pour éviter ce bouclage, en vérifiant que les itérées sont toujours supérieures à la première solution calculée.*

Remarque II - Tests d'arrêt de Convergence

Un test souvent employé utilise la distance relative (Cf. Par exemple, § 1) ou une formule similaire et contrôle que la distance entre deux itérées successives est inférieure à un nombre imposé et petit. Il consiste donc à vérifier que les itérations successives se stabilisent vers une valeur que l'on prendra comme solution calculée. MAIS CE TEST EST INSUFFISANT *car il ne peut tenir compte de phénomènes telles que :*

- Erreur sur la monotonie lorsque $x_{i+1}$ est inférieur à $x_i$
- Erreur sur le calcul du Hessien qui devient négatif par suite des erreurs de calcul
- Erreur sur la convergence (ou "saut involontaire") lorsque le calcul cesse de converger vers une solution $r_j$ pour converger vers la solution suivante $r_{j+1}$.

*Ce test suppose d'autre part, qu'il est possible de calculer toutes les solutions avec une précision imposée.* Mais par suite des erreurs de calcul et de la précision finie de l'arithmétique utilisée, toute solution $r_j$ est inclus dans une zone d'incertitude (l'ensemble des réels indiscernables de la solution $r_j$), donc il est important de préciser la dimension. LE TEST QUE NOUS AVONS UTILISE ETUDIE L'EVOLUTION DE LA DISTANCE RELATIVE DE DEUX ITEREES SUCCESSIVES ET DETERMINE SI L'ON S'APPROCHE OU S'ELOIGNE DE LA SOLUTION LA PLUS PROCHE. Les phénomènes signalés ci-dessus sont alors interpretés soit comme entrée dans la zone d'incertitude (et l'itérée précédente indique alors une borne inférieure) soit comme sortie (et la longueur du saut que nous allons définir ci-dessous précise alors la borne supérieure).

Remarque III - Saut d'une solution calculée

Le saut permet d'obtenir une nouvelle valeur initiale, supérieure à la solution $r_j$ de multiplicité m qui vient d'être calculée, pour converger vers la solution suivante $r_{j+m}$.

Le saut est calculé par l'une des formules ci-dessous, en utilisant les 3 ou 4 dernières itérées :

$$x_o = 3\, x_i - 2\, x_{i-2}$$

$$x_o = 3\, x_i - 2\, x_{i-3}$$

ou, lorsque la convergence a été trop lente, en utilisant un saut relatif :

$$x_o = x_i\ (1 + \varepsilon \text{ signe } (x_i))$$

ou un saut absolu, au voisinage de zéro :

$$x_o = x_i + \varepsilon$$

( $\varepsilon$ est un nombre petit, dépendant de l'arithmétique utilisée, par exemple $10^{-6}$)

Une fois le saut calculé par l'une de ces formules, on vérifie que l'on est sorti de la zone d'incertitude et dans le cas contraire, on utilise la formule suivante jusqu'à ce que l'on soit suffisamment éloigné de la solution calculée. Il est nécessaire, en effet, de ne pas effectuer un saut trop grand pour ne pas "oublier" de solutions que l'on ne peut généralement pas détecter dans ce balayage. Nous verrons cependant au Chapitre III (page 11) que la règle des signes de Sturm permet de lever cette difficulté lorsque l'on applique la méthode de Laguerre au calcul des valeurs propres d'une matrice tridiagonale symétrique ou symétrisable.

Remarque IV - Multiplicité d'une solution

Le rapport $\mu(x) = \dfrac{P'^2(x)}{P'^2(x) - P(x)\,P''(x)}$ converge vers la multiplicité $m(r_j)$ de la solution $r_j$ lorsque x tend vers $r_j$. (Pour alleger les notations, nous parlerons de la multiplicité m de la solution r du polynôme P lorsqu'il n'y a pas ambiguïté.)

$$P(x) = (x-r)^m\, Q(x) \qquad \text{avec} \qquad Q(r) \neq 0$$

$$P' = (x-r)^{m-1}\left[mQ + (x-r)\,Q'\right]$$

$$P'' = (x-r)^{m-2}\left[m(m-1)\,Q + 2m(x-r)\,Q' + (x-r)^2\,Q''\right]$$

$$\frac{P'^2}{P'^2 - PP''} = \frac{mQ^2 + 2m(x-r)\,Q\,Q'' + (x-r)^2\,Q'^2}{mQ^2 + (x-r)^2\,(Q'^2 - Q\,Q'')} \xrightarrow[x \to r]{} m$$

Rappelons que

$$p = \frac{P'}{P} = \sum_{k=1}^{n} \frac{1}{x - r_k}$$

$$q^2 = \frac{P'^2 - PP''}{P^2} = \sum_{k=1}^{n} \left(\frac{1}{x - r_k}\right)^2$$

Alors

$$\mu = \frac{p^2}{q^2} \xrightarrow[x \to r]{} m(r)$$

$$x \sim r_j \Longrightarrow \left(\sum_{k=1}^{n} \frac{1}{x - r_k}\right)^2 \sim m(r_j) \sum_{k=1}^{n} \left(\frac{1}{x - r_k}\right)^2$$

*La connaissance de la multiplicité permet de compter le nombre de solutions calculées et, à la fin du balayage, de constater si des solutions ont pu être oubliées, mais elle ne permet pas, comme la règle des signes de Sturm, de déterminer où elles ont été oubliées. Dans ce cas, on procèdera à une division du polynôme par* $\prod (x - r_k^*)$ *où* $r_k^*$ *sont les solutions calculées.* Une nouvelle application de la méthode par saut indiquera les solutions oubliées que l'on sait proche des solutions précédemment calculées ou ne déterminera aucune solution (on pourra alors préciser que ce sont des solutions complexes et en indiquer le nombre).

Nous n'avons pas utilisé ce rapport $\mu$ pour tenter d'améliorer la convergence de la méthode vers une solution multiple. Remarquons cependant que ce rapport tend vers la multiplicité dans le voisinage de la solution cherchée, qu'il est indéterminé pour des solutions multiples (il faut dans ce cas le calculer pour des valeurs extérieures à la zone d'incertitude) et que sa valeur peut être imprécise pour des solutions très voisines.

## CHAPITRE III

Application de la Méthode de Laguerre au calcul des valeurs propres d'une matrice tridiagonale.

Il est toujours possible de déterminer une matrice tridiagonale semblable à une matrice donnée quelconque (Théorème de Lanczos) et il existe, pour cela, de nombreuses méthodes dont quatre sont particulièrement connues : La Méthode de Givens, historiquement l'une des premières, transforme une matrice symétrique par un produit de matrices de rotations. Celle de Householder, valable, elle aussi pour les seules matrices symétriques, utilise comme matrices de transformation, des matrices de polynôme minimal du second degré. La méthode de Wilkinson, transforme la matrice donnée quelconque en la multipliant par différentes matrices de passage (matrice de permutation ou matrice à un paramètre $I + m E_{ij}$) et conduit normalement à une matrice quasi-triangulaire puis à une matrice tridiagonale; une variante (J.F. MAITRE) permet même d'obtenir directement une matrice tridiagonale en annulant tour à tour un $a_{ij}$ $(i > j + 1)$ et le terme symétrique $a_{ji}$.

La méthode de Lanczos, la plus générale, utilise une matrice globale de transformation dont chaque colonne i est déduite (par un polynôme de degré i) de la première colonne.

Ces méthodes de tridiagonalisation sont assez coûteuses (de l'ordre de $N^3$ multiplications, N étant l'ordre de la matrice) mais se justifient en considérant l'ensemble des propriétés remarquables que possèdent les matrices tridiagonales et la relative facilité de déterminer les valeurs propres de telles matrices.

Nous verrons en effet qu'il est aisé de calculer, en un point $x_i$, les valeurs du polynôme caractéristique $P(x_i)$ et de ses dérivées successives ($P'(x_i)$, $P''(x_i)$...) ou de déterminer les coefficients de ces polynômes. Ces deux possibilités conduisent alors à deux méthodes de calcul des valeurs propres : l'une, où le point de vue polynômial est prépondérant, calcule les coefficients puis les solutions du polynôme caractéristique comme nous l'avons déjà vu au chapitre II ; l'autre, où l'on conserve un point de vue matriciel, permet de mieux utiliser les propriétés d'une matrice tridiagonale et notamment la règle des signes de Sturm dans le cas d'une matrice symétrique.

3.1 Calcul du Polynôme caractéristique et de ses dérivés successives.

Pour une valeur x donnée, le polynôme caractéristique peut être déterminé ainsi que ses dérivées successives, soit par le schéma de Sturm que nous allons préciser, soit par le schéma de Horner appliqué à ses coefficients.

Pour normaliser le polynôme caractéristique P(x) de la matrice tridiagonale T, nous prenons, comme nous l'avons déjà fait au chapitre II, le coefficient du terme de plus haut degré égal à 1, ce qui revient à calculer le déterminant de (x I - T)

D'autre part, pour alléger les notations, nous poserons :

$$(1)\quad \begin{cases} a_i = T_{i,i} & i \in [1, N] \\ b_i = T_{i+1,i} & i \in [1, N-1] \\ c_i = T_{i, i+1} & i \in [1, N-1] \end{cases}$$

On ne considère ainsi que 3 N - 2 éléments de la matrice au lieu de $N^2$.

3.1.1 Dans ces conditions, la relation de récurrence

$$(2)\quad P_i = (x - a_i)\, P_{i-1} - b_{i-1} \times c_{i-1}\, P_{i-2}$$

complétée par :

$$(3)\quad P_o = 1 \qquad P_1 = x - a_1$$

définit le schéma de Sturm où $P_i$ représente la valeur du sous-déterminant formé par les i premières lignes et colonnes de la matrice(x I - T) . $P_i$ est un polynôme de degré i. $P_N = P(x)$ est le polynôme caractéristique cherché.

3.1.3. Il est intéressant de noter, en effet, que dans toutes les relations précédemment définies, les termes $b_i$, $c_i$ n'interviennent que par leur produit $b_i \times c_i$ que l'on réalisera une fois pour toute. Plus généralement, nous définissons une matrice tridiagonale *réduite* T', semblable à la matrice T par :

$$(10)\quad \begin{cases} a'_i = a_i \\ b'_i = b_i \times c_i \\ c'_i = 1 \end{cases}$$

(et nous parlerons dans ce cas d'une *réduite inférieure*) qui possède donc les mêmes valeurs propres. Le calcul de T' à partir de T est peu coûteux (N - 1 multiplications) et le nombre d'élérents à conserver en mémoire, si l'on ne s'intéresse qu'aux valeurs propres, passe de 3 N - 2 à 2 N - 1.

Notons cependant une difficulté (lorsque l'on utilise une arithmétique en virgule fixe ou à tableaux encadrés, analogues à ceux définies par Wilkinson [11] sous l'appellation de "Block Floating" et dont nous avons déjà fait mention au chapitre I) :

*$b_i$ et $c_i$ peuvent être petits mais non nuls et donner lieu à un produit $b_i \times c_i$ indiscernable de 0.*

Le passage de T à T' peut se représenter par une transformation $S^{-1}$ T S où S est la matrice diagonale définie (dans le cas d'une matrice réduite inférieure) par :

$$(11)\quad \begin{cases} s_{1,1} = 1 \\ s_{i,i} = \prod_{j=1}^{i-1} c_j & i \in [2, n] \end{cases}$$

On en déduit de suite les coefficients de $S^{-1}$ :

$$(12)\begin{cases} S^{-1}_{1,1} = 1 \\ S^{-1}_{i,i} = \prod_{j=1}^{i-1} C_j^{-1} \qquad i \in [2, n] \end{cases}$$

qui ne sont définis que lorsque tous les $C_j$ sont différents de 0.

Nous dirons alors qu'une matrice tridiagonale est *réductible* si $b_i \times c_i \neq 0 \quad \forall_i \in [1, n-1]$

Dans le cas contraire, le polynôme caractéristique se factorise et nous aurons à déterminer les valeurs (et vecteurs) propres de deux matrices (provenant de la partition) ayant un ordre moins élevé. La notion de matrice réduite peut s'étendre aux matrice T ayant des termes extradiagonaux nuls mais la transformation par la matrice S définie ci-dessus n'est évidemment pas applicable.

Pour obtenir les vecteurs propres de la matrice T, il faudra multiplier par S ainsi défini, ceux de la matrice réduite T' ce qui risque de perturber considérablement les dernières composantes des vecteurs propres par suite des erreurs probables de multiplication $\varepsilon_m$. En effet, en appelant S*, la matrice effectivement calculée , on a :

$$(13)\begin{cases} S^*_{11} = S_{11} \\ S^*_{22} = S_{22} \\ S^*_{33} \approx (1 + \varepsilon_m)\, S_{33} \\ \vdots \\ S^*_{ii} \approx (1 + (i-2)\, \varepsilon_m)\, S_{ii} \\ S^*_{NN} \approx (1 + (N-2)\, \varepsilon_m)\, S_{NN} \end{cases}$$

Cette transformation n'est donc pas à utiliser lors

l'on désire aussi calculer les vecteurs propres.

3.1.4. comparaison des temps de calcul du polynôme caractéristique d'une matrice tridiagonale d'ordre N pour n valeurs de x, par le schéma de Sturm ou de Horner.

Rappelons qu'un horner correspond au temps utilisé pour le calcul d'un polynôme de degré N par le schéma de Horner et équivaut au temps de calcul de (N-1) multiplications et N additions.

| | Schéma de Sturm | Schéma de Horner |
|---|---|---|
| Matrice réduite | N-1 multiplications | N-1 multiplications |
| Coefficients | | $\frac{(N-1)\ (N-2)}{}$ multiplications<br>$\frac{(N-1)\ (N-2)}{}$ additions |
| Schéma | n(7N-17) multiplications<br>n(4N-5) additions | 3n (N-1) multiplications<br>3n N additions |
| Total | 7n + 1 horners | 3n + N/2 horners |

3.2. Propriétés des matrices symétrisables.

3.2.1. Définition

Nous dirons qu'une matrice est symétrisable, s'il existe une matrice symétrique qui lui est semblable. Elle possède alors des valeurs propres toutes réelles. Cette notion est particulièrement intéressante pour les matrices tridiagonales que nous considérons ici.

3.2.2. Propriétés des matrices symétrisables.

Théorème I : Soit, en effet, une matrice tridiagonale T et $b_i$, $c_i$ ses éléments extradiagonaux inférieurs et supérieurs. Alors :

(14) $$\boxed{\forall i \in [1, N-1] \quad : \quad b_i \, c_i > 0 \Longrightarrow T \text{ symétrisable}}$$

Considérons la matrice diagonale S définie par :

(15) $$\begin{cases} S_{11} = 1 & S_{ii} = \sqrt{\prod_{j=1}^{i-1} (c_j \, b_j^{-1})} \\ S_{11}^{-1} = 1 & S_{ii}^{-1} = \sqrt{\prod_{j=1}^{i-1} (b_j \, c_j^{-1})} \end{cases}$$

La similitude $S \, T \, S^{-1}$ transforme la matrice T en une matrice semblable T' ayant pour éléments (avec les notations habituelles) :

(16) $$\begin{cases} a'_i = a_i \\ b'_i = c'_i = \sqrt{b_i \, c_i} \end{cases} \qquad i \in [1, N-1]$$

Les vecteurs propres sont modifiés par la similitude. Si l'on désire calculer les vecteurs propres de T, connaissant ceux de T', il faudra multiplier ceux-ci par la matrice de passage S. Malgré la présence dans le calcul de S, de racines carrées qui améliorent la précision des calculs, ce produit risque de perturber considérablement les dernières composantes comme le montre une analyse d'erreur, analogue à celle indiquée pour les matrices réduites.

Nous upposons que $\varepsilon_m$ et $\varepsilon_d$ représentent l'erreur probable de multiplication et de division et q e la racine carrée diminue l'erreur de moitié. Alors :

$$(17)\quad \begin{cases} S^*_{11} = S_{11} \\ S^*_{22} \approx (1 + \frac{\varepsilon_d}{2}) S_{22} \\ S^*_{33} \approx (1 + \varepsilon_d + \frac{\varepsilon_m}{2}) S_{33} \\ S^*_{ii} \approx (1 + \frac{(i-1)\varepsilon_d + (i-2)\varepsilon_m}{2}) S_{ii} \\ S^*_{NN} \approx (1 + \frac{(N-1)\varepsilon_d + (N-2)\varepsilon_m}{2}) S_{NN} \end{cases}$$

Il ne peut donc être conseillé de calculer ainsi les vecteurs propres.
Remarquons aussi que toute méthode de détermination des valeurs propres d'une matrice tridiagonale symétrique utilisant le polynôme caractéristique d'une manière implicite (schéma de Sturm) ou explicite (schéma de Horner appliqué aux coefficients du polynôme) nécessite le calcul de $b'_i \times c'_i = b_i \times c_i$. Cette transformation est alors inutile.

La réciproque du Théorème I est évidemment fausse. Par exemple, une matrice tridiagonale symétrique donc symétrisable peut avoir des éléments $b_i\ c_i$ nuls. On peut cependant affirmer que :

Théorème II : Pour une matrice tridiagonale T d'ordre N

$$(18)\quad \boxed{\forall i \in [1, N-1] : b_i\ c_i \geq 0 \Longrightarrow \text{T a des valeurs propres réelles}}$$

La matrice T se décompose en effet en sous-matrices symétrisables qui ont leurs valeurs propres réelles.

Théorème III :

Soit une matrice tridiagonale T d'ordre N

(19) $$\boxed{T \text{ a ses valeurs propres réelles} \Longrightarrow 2 \sum_{i=1}^{N-1} b_i c_i \geqslant - \sum a_i^2}$$

En effet : Si une matrice T a ses valeurs propres réelles, la matrice $T^2$ a des valeurs propres toutes positives et la somme de ses termes diagonaux est positive.

Les termes diagonaux de la matrice $T^2$ sont

(20) $$\begin{cases} \tau_{11} = a_1^2 + b_1 c_1 \\ \vdots \\ \tau_{ii} = b_{i-1} c_{i-1} + a_i^2 + b_i c_i \qquad i \in [2, N-1] \\ \vdots \\ \tau_{NN} = a_N^2 + b_{N-1} c_{N-1} \end{cases}$$

(21) $$s^2 = \sum_{i=1}^{N} a_i^2 + 2 \sum_{i=1}^{N-1} b_i c_i \geqslant 0$$

On en déduit aussi :

(22) $$\boxed{\sum_{i=1}^{N-1} b_i c_i < - \frac{1}{2} \sum_{i=1}^{N} a_i^2 \Longrightarrow T \text{ a des valeurs propres complexes.}}$$

Suites de Sturm

3.2.3. Il est intéressant de remarquer que la règle des signes de Sturm définie pour des matrices symétriques T' peut s'étendre aux matrices symétrisables T (tridiagonales).

Nous avons vu en effet que $b'_i \times c'_i = b_i \times c_i \quad \forall i \in [1, N-1]$

Les suites $P_i$ de Sturm correspondantes sont alors identiques car les éléments extradiagonaux n'interviennent que par leur produit. Pour des matrices symétrisables, il est donc utile de conserver un point de vue matriciel utilisant ainsi la règle des signes de Sturm que nous allons rappeler [12]

POUR UNE VALEUR x DONNEE, LE NOMBRE DES VARIATIONS DE SIGNES DE LA SUITE $P_i$ (FORMEE POUR CALCULER LA VALEUR DU POLYNOME CARACTERISTIQUE) EST EGAL AU NOMBRE DE VALEURS PROPRES DE T SUPERIEURES A x.

Cette règle permet donc de vérifier le choix correct d'une valeur initiale, de déterminer la multiplicité d'une valeur propre et d'indiquer si le saut effectué après le calcul d'une valeur n'est pas trop grand (valeurs propres "oubliées"). Son utilisation conduit à l'algorithme suivant :
Un balayage par la méthode de Laguerre localise les valeurs propres dans de très petits intervalles. L'ordre de multiplicité est alors obtenu par le comptage du nombre de variations (ou par le calcul de $M(x_i)$ comme nous l'avons vu au chapitre II, lorsque l'on désire conserver un point de vue polynomial ou lorsque la matrice n'est pas symétrisable.

On a donc totalement résolu en un seul balayage le problème suivant dont nous avons déjà remarqué qu'il avait seul un sens :

*Trouver des zones de taille minimale, si possible les zones d'indiscernabilité en fonction des incertitudes, contenant les solutions et estimer le nombre de solutions intérieures à chaque zone*

L'algorithme de balayage utilise :

- Les tests d'arrêts de convergence et le saut définis au chapitre II
- Une localisation des valeurs propres selon l'une ou l'autre des possibilités suivantes, qui s'appuient respectivement sur les théorèmes de Laguerre ou de Gerschgorin.

### 3.3. Localisation des valeurs propres

Il est important en effet de bien localiser les valeurs propres pour limiter le temps global des calculs, et pour éviter les difficultés apportées par l'infini-machine.

#### 3.3.1. Localisation par la méthode de Laguerre

Nous avons vu au chapitre II que itérer l'infini par la méthode de Laguerre conduisait à deux valeurs encadrant les solutions du polynôme. Le calcul de ces valeurs nécessite la connaissance des coefficients $C_1$ et $C_2$ du polynôme caractéristique

$$P(x) = x^n + C_1 x^{n-1} + C_2 x^{n-2} + \text{-------} C_n \qquad (23)$$

que l'on sait aisément déterminer pour une matrice tridiagonale.

En appelant $\lambda_k$ les valeurs propres de T (et par suite les solutions de son polynôme caractéristique)

$$(24)\quad C_1 = -\sum_{k=1}^{N} \lambda_k = -\sum_{i=1}^{N} a_i \quad \text{(trace de la matrice)}$$

$T^2$ a pour valeur propre $\lambda_k^2$ et pour termes diagonaux :

$$(25)\quad \tau_{ii} = b_{i-1}\ c_{i-1} + a_i^2 + b_i\ \alpha_i$$

ou pour une matrice réduite

$$(26)\quad \tau_{ii} = b_{i-1} + a_i^2 + b_i$$

Comme $C_2 = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \lambda_i\ \lambda_j$

$$= \sum_{\substack{i=1 \\ i \neq j}}^{N} \sum_{j=1}^{N} \lambda_i\ \lambda_j$$

$$(27)\quad C_2 = \frac{1}{2}\left(\sum_{i=1}^{N} \lambda_i\right)^2 - \frac{1}{2}\sum_{i=1}^{N} \lambda_i^2$$

$$= \frac{1}{2}\left[\left(\sum_{i=1}^{N} a_i\right)^2 - \sum_{i=1}^{N} (a_i^2 + 2\ b_i)\right]$$

où l'on aura posé $b_o = b_n = 0$ et remarqué que

$$\sum_{i=1}^{N} (b_i + b_{i-1}) = 2 \sum_{i=1}^{N} b_i$$

Les bornes cherchées sont alors données par ($\upsilon^2 = 1$) :

$$(28)\quad \mathcal{L}_n(\infty) \sim \frac{-C_1 + \upsilon\sqrt{(N-1)^2\ C_1^2 - 2\ N\ (N-1) C_2}}{N}$$

Coût par Laguerre

| | Additions | Multiplications | Divisions | Racine Carrée |
|---|---|---|---|---|
| $C_1$ | N | | | |
| $C_2$ | 2N+1 | 2N + 1 | 1 | |
| $\mathcal{L}_n(\infty)$ | 3 | 5 | 2 | 1 |
| Total | 3 N + 4 | 2 N + 6 | 3 | 1 |

3.3.2 Localisation par les cercles de Gerschgorin

Il peut être intéressant de remplacer cette méthode de localisation dérivée de la méthode de Laguerre par une autre utilisant les cercles de Gerschgorin dont nous allons brièvement rappeler la définition.

Remarquons cependant que la procédure Algol Bornes (en Annexe à ce chapitre) indique une localisation des valeurs propres en calculant l'intersection des intervalles définis par ces deux méthodes.

Pour une matrice A quelconque de valeurs propres $\lambda_k$, le théorème de Hadamard permet d'affirmer que :

(29)
$$\forall \lambda_k \exists r : \quad |a_{rr} - \lambda| \leq \sum_{i=1}^{N} |a_{ir}| = \rho_r$$

Dans le plan complexe, l'ensemble $\Lambda$ des valeurs propres est donc contenu dans la réunion des disques de centre $a_{rr}$ et de rayon $\rho_r$ qui forme alors un domaine de localisation.

(30)
$$\Lambda \subset \bigcup_r \left\{ a_{rr}, \sum_{\substack{i=1 \\ i \neq r}}^{N} |a_{ir}| \right\}.$$

On obtient un autre domaine de localisation en considérant les "cercles" de centre $a_{rr}$ et de rayon

$$\rho'_r = \sum_{i \neq r} |a_{ri}| .$$

D'où

$$(31)\quad \Lambda \subset \bigcup_r \{a_{rr},\ \sum_{i\neq r} |a_{ir}|\} \cap \bigcup_r \{a_{rr}\ \sum_{i\neq r} |a_{ri}|\}.$$

Plus généralement, l'intersection d'un nombre quelconque de domainesde localisation (contenant toutes les valeurs propres) est lui-même un domaine de localisation.

Pour localiser les valeurs propres des matrices à coefficients réels et à valeurs propres réelles, on prendra un intervalle de localisation défini par la trace de ce domaine sur l'axe réel.

Remarque - Les intervalles de localisation définis pour une matrice tridiagonale et sa matrice réduite ne sont pas identiques car le domaine de localisation n'est pas une fonction des produits $b_i \times c_i$ seulement. On peut donc calculer ces deux intervalles et en prendre l'intersection (à défaut de minimiser (cf. travaux de Bauer, par ex) les domaines correspondants à tout $S^{-1}.A.S$)

Coût par Gerschgorin

| | Additions |
|---|---|
| Total | 8 N - 8 |

## 3.4. Calcul des vecteurs propres d'une matrice tridiagonale

Une relation existe entre les composantes des vecteurs propres d'une matrice tridiagonale T indecomposable ($\forall_i : b_i \times c_i \neq 0$) et les suites de polynômes $P_i$ et $Q_i$ que nous avons définies. Elle a été étudiée notamment par Nolin [13] ou , dans un cas particulier, par J.H. Wilkinson [14].
Notre exposé suivra d'ailleurs de très près les travaux précités.

Rappelons que les deux suites $P_i$ et $Q_i$ donnent la valeur du polynôme caractéristique $P(x)$ de la matrice $(xI - T)$ pour une valeur x quelconque :

(33) $P(x) = P_N(x) = Q_N(x)$.

Soient $\lambda$ une valeur propre de T et V le vecteur propre correspondant :

(34) $V = |v_1, v_2, \text{------}, v_n|$
avec $v_1 = 1$.

V et $\lambda$ vérifient alors le système

(35) $(\lambda I - T) V = 0$.

Considérons le système linéaire formé par les $(N - 1)$ premières équations de (35)

$$(36)\quad \begin{cases} (\lambda - a_1) v_1 - c_1 v_2 = 0 \\ -b_1 v_1 + (\lambda - a_2) v_2 - c_2 v_3 = 0 \\ \vdots \\ -b_{i-1} v_{i-1} + (\lambda - a_i) v_i - c_i v_{i+1} = 0 \\ \vdots \\ -b_{N-2} v_{N-2} + (\lambda - a_{N-1}) v_{N-1} - c_{N-1} v_N = 0 \end{cases}$$

En résolvant ce système, nous obtenons :

$$(37)\quad \begin{cases} v_1 = 1 = P_o(\lambda) \\ v_2 = \dfrac{\lambda - a_1}{c_1} = \dfrac{P_1(\lambda)}{c_1} \\ v_3 = \dfrac{(\lambda - a_2) P_1 - b_1 c_1 P_o}{c_1 c_2} = \dfrac{P_2(\lambda)}{c_1 c_2} \\ \vdots \\ v_i = \dfrac{(\lambda - a_{i-1}) P_{i-2} - b_{i-2} c_{i-2} P_{i-3}}{c_1 c_2 \text{------} c_{i-1}} = \dfrac{P_{i-1}(\lambda)}{c_1 c_2 \text{---} c_{i-1}} \end{cases}$$

$$v_{N-1} = \frac{(\lambda - a_{N-2})\, P_{N-3} - b_{N-3}\, c_{N-3}\, P_{N-4}}{c_1\ c_2 \cdots c_{N-2}} = \frac{P_{N-2}(\lambda)}{c_1\, c_2 \cdots c_{N-2}}$$

$$v_N = \frac{(\lambda - a_{N-1})\, P_{N-2} - b_{N-2}\, c_{N-2}\, P_{N-3}}{c_1\ c_2 \cdots c_{N-1}} = \frac{P_{N-1}(\lambda)}{c_1\, c_2 \cdots c_{N-1}}$$

ce qui s'écrit, en posant $c_o = 1$

$$(38) \qquad \boxed{v_i = \frac{P_{i-1}(\lambda)}{c_o c_1\ c_2 \cdots c_{i-1}} \qquad \forall i \in [1, N]}$$

Le premier membre $(-b_{N-1}\, v_{N-1} + (x - a_N)\, v_N)$ de la dernière équation du système (35) s'écrit alors :

$$(39) \quad -b_{N-1}\, v_{N-1} + (x - a_N)\, v_N = \frac{(x-a_N)\, P_{N-1}(x) - b_{N-1} c_{N-1}\, P_{N-2}(x)}{c_1\ c_2 \cdots c_{N-1}}$$

$$= \frac{P_N(x)}{c_1\ c_2 \cdots c_{N-1}}$$

quantité qui est bien nulle pour la valeur $\lambda$.

On peut aussi déterminer les coefficients du vecteur propre $U$ à gauche correspondant à $\lambda$, en considérant :

$$(40) \quad U\,(\lambda I - T) = 0$$

Nous résolvons les $(N - 1)$ premières équations du système (40) et posons $b_o = 1$.

$$(41) \qquad \boxed{U_i = \frac{P_{i-1}(\lambda)}{b_o\, b_1 \cdots b_{i-1}} \qquad \forall i \in [1, N]}$$

En introduisant la suite des $Q_i$, on obtient d'autres expressions des coefficients (notés $u'_i$ et $v'_i$) des vecteurs propres U et V, par résolution des (N-1) dernières équations des systèmes (40) et (35). ($b_N = c_N = 1$)

$$(42)\quad u'_i = \frac{Q_{N-i}(\lambda)}{c_i \cdots c_{N-1}\, c_N} \qquad \forall_i \in [1, N]$$

$$(43)\quad v'_i = \frac{Q_{N-i}(\lambda)}{b_i \cdots b_{N-1}\, b_N} \qquad \forall_i \in [1, N]$$

Notons que, dans ce cas, les coefficients $u'_N$ et $v'_N$ sont pris égaux à 1.

Remarque I - Toutes ces formules supposent que les produits $b_i$, $c_i$ ne sont pas nuls (ces éléments apparaissent en dénominateur) Lorsque l'un des éléments extradiagonaux est nul (par exemple $c_i = 0$, $b_i \neq 0$) la matrice se décompose de la manière suivante :

$$(44)\quad T = \begin{bmatrix} A & 0 \\ B & D \end{bmatrix}$$

où A et D sont des matrices carrées d'ordre i et N-i ; B est une matrice ayant tous ses éléments nuls sauf le terme $t_{i+1,i} = b_i \neq 0$. Les valeurs propres de la matrice T sont celles des matrices A et D.

$$(45)\quad (\lambda I - T)\, V = \begin{bmatrix} \lambda I_1 - A & 0 \\ -B & (\lambda I_2 - D) \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = 0$$

Soit $\lambda$ et $V_2$ une valeur propre et le vecteur propre correspondant de D ; alors le vecteur propre de T correspondant à $\lambda$ est formé de $V_1 = 0$ et $V_2$ vecteur propre de D.

Si $\lambda$ est valeur propre de A seulement, on prendra pour $V_1$ le vecteur propre correspondant de A et pour $V_2$ le vecteur sinsi défini :

(46) $V_2 = (\lambda I_2 - D)^{-1} E$

où

(47) $E = B V_1$

est un vecteur colonne dont la première composante seule est non nulle. On ne calculera donc que la première colonne de l'inverse de $(\lambda I_2 - D)$.

Lorsque l'un des éléments extradiagonaux et sont symétrique sont nuls (par exemple $b_i = c_i = 0$), la matrice se décompose simplement :

(48) $T = \begin{bmatrix} A & 0 \\ 0 & D \end{bmatrix}$

et les vecteurs propres de T sont ceux de A et D complétés par des zéros.

(49) $\lambda$ valeur propre de $A \Rightarrow \begin{bmatrix} \lambda I_1 - A & 0 \\ 0 & \lambda I_2 - D \end{bmatrix} \begin{bmatrix} V_1 \\ 0 \end{bmatrix} \neq 0$

(50) $\lambda$ valeur propre de $D \Rightarrow \begin{bmatrix} \lambda I_1 - A & 0 \\ 0 & \lambda I_2 - D \end{bmatrix} \begin{bmatrix} 0 \\ V_2 \end{bmatrix} = 0$

Remarque II : Supposons maintenant tous les $b_i$ $c_i$ non nuls. Le coefficient de proportionnalité entre les vecteurs définis par les coefficients $v_i$ et $v'_i$ est

(51) $$\rho = \frac{Q_{N-1}(\lambda)}{c_1 c_2 \cdots c_N}$$

qui ne peut être nul.

En effet rappelons que, pour $x = \lambda$

(52) $Q_i(\lambda) = (\lambda - a_{N-i+1}) Q_{i-1} - b_{N-i+1} \times c_{N-i+1} Q_{i-2}$

(53) $Q_N(\lambda) = P(\lambda) = 0$

Alors :

$$(54)\quad Q_N(\lambda) = Q_{N-1}(\lambda) = 0 \Longrightarrow Q_{N-2}(\lambda) = 0 \cdots\cdots\Rightarrow Q_o = 0$$

ce qui est impossible car par hypothèse $Q_o = 1$.

On en déduit que lorsque tous les $b_i$, $c_i$ sont différents de zéro, on a bien déterminé la direction propre associée à la valeur propre $\lambda$.

Remarque III - Lorsque l'on calcule les vecteurs propres d'une matrice réduite supérieure, les coefficients sont donnés plus simplement par :

$$(55)\quad v_i = P_{i-1}(\lambda) \quad \forall i \in [1, N] \quad \text{car } c_i = 1 \quad \forall_i$$

On obtient donc sans calcul supplémentaire les vecteurs propres de telles matrices. La précision est donc celle de la suite des $P_i$. Il est est tout différemment pour une matrice tridiagonale quelconque car la division par le produit des $c_i$ risque de perturber considérablement les dernières composantes comme nous l'avons vu en présentant la notion de matrice réduite. Il est en fait équivalent de calculer les vecteurs propres d'une matrice quelconque par la méthode que nous venons d'exposer ou de les déduire de la matrice réduite semblable en les multipliant par la matrice de passage S, car le calcul perturbateur (division par le produit des $c_i$) intervient dans les deux cas. Nous utilisons donc systématiquement la notion de matrice réduite qui présente des avantages intéressants pour le calcul des valeurs propres.

Le programme réalisé n'utilise pas la possibilité de stabiliser certains calculs en effectuant des homotheties ou des translations qui, sans grande influence sur les calculs et leur coût, influencent par contre les tests d'arrêt du type "distance relative".

Connaissant une valeur propre et le vecteur propre correspondant, on peut utiliser, pour diminuer le temps global de calcul, une méthode de deflation (ce qui réduit l'ordre de la matrice) et notamment celle proposée par Nolin [15] pour les matrices tridiagonales.
Cette méthode (ou sa variante qui conserve la forme de matrice réduite), modifie profondément les vecteurs propres et perturbent le calcul des valeurs propres. Elles ne sont pas utilisées dans ce programme car on calcule toutes les valeurs propres par un seul balayage sans division successive (point de vue polynomial) ou deflation (point de vue matriciel)

## 3.5. Comparaison des points de vue polynomial et matriciel

| | Point de vue Polynomial | Point de vue Matriciel |
|---|---|---|
| Coût de k itérations de Laguerre | 3 k + N/2 horners beaucoup plus économique | 7 k + 1 horners |
| Localisation | Par la formule de Laguerre | Même possibilité + Cercles de Gerschgorin |
| Multiplicité | $M_i = \lim_{x \to \lambda_i} \frac{P'^2(x)}{P'^2(x) - P(x)P''(x)}$ | Même possibilité + règle des signes de Sturm pour les matrices symétriques Gros avantage, encore que la règle de Sturm donne lieu à des résultats instables. |
| Calcul des solutions | Rapidités de convergence —— | identiques |
| Test d'Arrêt | identique —— | —— |
| Calcul des vecteurs propres | Impossible directement | facile(mais instable) directement |
| Amélioration des solutions en amas par 2ème balayage | Possible après division | Possible après déflation (Instable) |

En conclusion on adoptera le point de vue polynomial (à cause de sa grande économie) sauf dans les cas suivants :

a) La matrice est symétrisable ($\forall_i : b_i \times c_i > 0$)
b) On désire aussi les vecteurs propres?

Remarque - Les calculs par la seconde méthode (Sturm) possèdent une signification matricielle que n'ont pas ceux auxquels conduit la première méthode (Horner). Notamment, les coefficients du polynôme caractéristique sont fort incertains alors que la méthode par Sturm n'oblige pas à passer par eux ; mais ceci ne signigie pas nécessairement que la méthode matricielle soit plus stable que la méthode polynomiale.

# CHAPITRE IV

## LISTE DE MATRICES-TEST

Les programmes décrits ci-dessus calculent les valeurs propres d'une matrice mise au préalable sous la forme tridiagonale. Les matrices test présentées dans ce chapitre seront donc de cette forme et peuvent être réparties en deux catégories.

1°/ Celles que l'on peut définir par une formule mathématique et dont on peut calculer analytiquement les valeurs propres. Ces matrices notées ($T_1$, $T_2$, ------, $T_9$) sont alors précieuses pour éprouver un programme de calcul.

2°/ Celles que l'on peut aussi définir par une formule mais dont les valeurs propres ne sont pas connues exactement. Nous en indiquons trois ($T_{10}$, $T_{11}$, $T_{12}$) qui ont déjà donné lieu à plusieurs publications.

Cette liste évidemment incomplète, ne comporte pas en particulier de matrices définies numériquement car on ne peut assurer alors une définition exacte en machine. Elle ne contient que des matrices non nécessairement symétriques mais à valeurs propresù toutes réelles.

On donne en annexe le calcul (analytique) des valeurs propres de la matrice de Wolstenholme et de quelques matrices qui se déduisent de celle-ci.

4.1 La matrice symétrique d'ordre N.

$$t_{i,i+1} = t_{i+1,i} = 1 \qquad \forall_i \in [1, N]$$

$$t_{i,j} = 0 \qquad \forall_i \neq j + 1 \text{ ou } j - 1$$

est connue sous le nom de matrice de Wolstenholme.
Par exemple, pour N = 5

$$T_1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

D.E. Rutherford [16] et Lee Parker [17] ont donné les valeurs propres $\lambda_k$ de cette matrice :

$$\boxed{\lambda_k = 2 \cos \frac{k \pi}{N + 1} \qquad k \in [1, N]}$$

résultat que l'on retrouve aisément en effectuant le changement de variable $\lambda = u + u^{-1}$. Connaissant les valeurs propres de cette matrice $T_1$, il est possible de calculer analytiquement celles de matrices s'en déduisant par translation ou par d'autres processus simples.

4.2. La matrice $T_1(a)$ obtenue par translation d'une valeur (a)

$$t_{i,i} = a \qquad i \in [1, N]$$

$$t_{i,i+1} = t_{i+1,i} = 1 \qquad i \in [1, N-1]$$

$$t_{i,j} = 0 \qquad |i-j| > 1$$

a pour valeurs propres :

$$\lambda_k = a + 2 \cos \frac{k \pi}{N + 1} \qquad k \in [1, N]$$

Remarque I - Pour N impair, cette matrice possède une valeur propre égale à (a)

Elle intervient notamment en géodesie (a = 3) ou dans la théorie des équations différentielles sous la forme;(N=5 par exemple et a = -2)

$$T_2 = \begin{bmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix}$$

donc les valeurs propres sont

$$\lambda_k = 4 \sin^2 \frac{k \pi}{2 (N + 1)} \qquad k \in [1, N]$$

Remarque II - Les valeurs propres de $T_2$ vérifient $\boxed{\Pi \lambda_k = N + 1}$

Remarque III - La matrice $T_2$ a une inverse symétrique connue analytiquement

$$a_{ij} = i (N + 1 - j) / (N + 1) \qquad i \leq j$$

$$a_{ij} = a_{ji} \qquad i > j$$

Par exemple (N = 5)

$$T_2^{-1} = \frac{1}{6} \begin{bmatrix} 5 & 4 & 3 & 2 & 1 \\ 4 & 8 & 6 & 4 & 2 \\ 3 & 6 & 9 & 6 & 3 \\ 2 & 4 & 6 & 8 & 4 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix}$$

4.3. Pour une valeur (a) très grande la matrice $T_1(a)$ a toutes ses valeurs propres différentes mais relativement très proches : elle présente donc un grand intérêt lorsque l'on désire éprouver un indicateur de multiplicité. (Nous noterons $T_3$ la matrice obtenue avec a = 1000).

4.4. On peut déduire d'autres matrices-test de la matrice de Wolstenholme notamment la matrice $T_4$ (A, B, C)

$$\left.\begin{aligned} t_{i,i} &= A \\ t_{i+1,i} &= B \\ t_{i,i+1} &= C \end{aligned}\right\} \quad BC > 0$$

dont les valeurs propres sont :

$$\lambda_k = A - 2\sqrt{BC}\cos\frac{k\pi}{N+1} \qquad k \in [1, N]$$

En choisissant les valeurs de A, B, C (Par exemple A = 6, B = 4, C = 9) on peut étaler les valeurs propres sur l'axe réel.

$$T_4 = \begin{bmatrix} 6 & 9 & 0 & 0 & 0 \\ 4 & 6 & 9 & 0 & 0 \\ 0 & 4 & 6 & 9 & 0 \\ 0 & 0 & 4 & 6 & 9 \\ 0 & 0 & 0 & 4 & 6 \end{bmatrix}$$

4.5. Une matrice de Wolstenholme de type 2 est obtenue en prenant $t_{N,N} = -1$. Cette matrice $T_5$ possède des valeurs propres connues :

$$\lambda_k = 2\cos\frac{2k\pi}{2n+1} \qquad k \in [1, N]$$

et une inverse symétrique connue.

$$a_{ij} = i \qquad i \leqslant j$$

$$a_{ij} = a_{ji} \qquad i > j$$

Par exemple pour N = 5

$$T_5 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix} \qquad T_5^{-1} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 & 3 \\ 1 & 2 & 3 & 4 & 4 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix}$$

4.6. Une matrice de Woltenholme de type 3 s'obtient en modifiant les termes $t_{1,1}$ et $t_{N,N}$

Notons, par exemple, la matrice $T_6$ obtenue en imposant $t_{1,1} = -1$ et $t_{N,N} = -1$. Elle possède les valeurs propres

$$\lambda_k = 2 \cos \frac{k \pi}{N} \qquad k \in [1, N]$$

et en particulier la valeur propre zéro pour N pair.

$$T_6 = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

4.7. La matrice symétrique définie par

$$a_{ij} = 2 \text{ Min } (i, j) - 1$$

est connue sous le nom de matrice de Rutherford [18].

Elle est indiquée dans la liste de M. Newman et J. Todd [19]

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 3 & 3 & 3 \\ 1 & 3 & 5 & 5 & 5 \\ 1 & 3 & 5 & 7 & 7 \\ 1 & 3 & 5 & 7 & 9 \end{bmatrix}$$

Elle possède une inverse tridiagonale (notée $T_7$)

$$a_{i,i} = 1 \qquad i \in [2, N-1]$$

$$a_{1,1} = 3/2$$

$$a_{N;N} = 1/2$$

$$a_{i,i+1} = a_{i+1,i} = -1/2 \qquad i \in [1, N-1]$$

$$a_{i,j} = 0 \qquad i \neq j\ ,\ i \neq j+1$$

qui se déduit d'une matrice de Wolstenholme de type 3

(où $t_{1,1} = 1$, $t_{N,N} = -1$)

$$T_7 = \frac{1}{2} \begin{bmatrix} 3 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

Cette remarque permet le calcul analytique de ses valeurs propres :

$$\lambda_k = 1 + \cos \frac{(2k-1)\,\pi}{2N} \qquad k \in [1, N]$$

4.8. Les polynômes de Tchebichef obéissant aux relations de récurrence

$T_{n+1} = 2 x T_n - T_{n-1}$ avec $T_o = 1$ , $T_1 = x$

ont pour solutions, les valeurs propres de la matrice $T_8$.

$$T_8 = \frac{1}{2} \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 2 & 0 \end{bmatrix} \quad \text{pour } N = 5$$

$$\lambda_k = \cos \frac{(2k-1)\Pi}{2N} \qquad k \in [1, N]$$

c'est-à-dire, les mêmes valeurs propres que la matrice $T_7$ à une translation près.

4.9 P.J. Eberlein [20] indique une matrice à deux paramètres (l'ordre et un paramètre arbitraire s) dont on connait analytiquement les valeurs et vecteurs propres. Cette matrice intervient dans la théorie chimique des gazs et peut être utilisée comme matrice-test spécialement pour les valeurs négatives du paramètre s. L'auteur indique que les valeurs propres sont particulièrement mal conditionnées pour $s < -N$. Cette matrice, notée $T_9$ est définie par :

$$a_{ii} = -(2i-1)N - (i-1)s + (2i^2 - 2i + 1)$$

$$a_{i,i+1} = i(N + s - i)$$

$$a_{i+1,i} = i(N - i)$$

$$a_{ij} = 0 \qquad \forall_{i,j} \quad |i-j| > 1$$

On peut aussi définir $a_{ii}$ par la forme plus économique :

$$a_{ii} = 2i^2 - i(2N + s + 2) + (N + s + 1)$$

Les valeurs propres sont toujours réelles, indépendantes de l'ordre N de la matrice. La valeur propre zéro est toujours présente.

$$\lambda_k = -(k-1)(s+k) \quad k \in [1, N]$$

La matrice $T_9$ est symétrisable si $s \geq -1$. En effet

$$\forall_i \in [1, N-1] \quad s \geq -1 \Longrightarrow a_{i,i+1} \times a_{i+1,i} = i^2 (N+s-i)(N-i) \geq 0$$

Dans ce cas toutes les valeurs propres sont simples, négatives ou nulles.

Cette matrice $T_9$ a des valeurs propres doubles si :
s entier et $-2N + 1 \leq s \leq -1$

En effet : soient $\lambda_j = \lambda_k$ , $j \neq k$

$(j-1)(s+j) = (k-1)(s+k)$

$(j-k)(j+k+s-1) = 0$

$\Longrightarrow j + k + s - 1 = 0$

$$\left.\begin{array}{l} \forall s \text{ entier} : -2n + 1 \leq s \leq -1 \\ \forall_{j,k} \in [1, N] : \quad j+k+s-1 = 0 \end{array}\right\} \Longrightarrow \lambda_j = \lambda_k$$

Par exemple : $s = -6$, $N \geq 6$

$\lambda_1 = \lambda_6 = 0$
$\lambda_2 = \lambda_5 = 4$
$\lambda_3 = \lambda_4 = 6$

Les valeurs propres doubles sont toujours positives ou nulles.

Remarque : N Pair ($N = 2\mu$), $s = -2\mu$ $\Longrightarrow$ Toutes les valeurs propres sont double

N impair ($N = 2\mu+1$), $s = -2\mu - 1$ $\Longrightarrow$ $\forall_i \quad 0 \leq \lambda_i \leq \lambda_{\mu+1}$

$\forall_i \neq \mu + 1$ $\lambda_i$ est valeur propre double

$s < -n \Longrightarrow \forall_i \quad \lambda_i \geq 0$

4;10 Une matrice introduite par D.H. Lehmer, cité dans la liste de Newman et Todd [19] est définie par

$$a_{ij} = i / j \qquad i \leq j$$

$$a_{ij} = j / i \qquad i > j$$

$$A = \begin{bmatrix} 1 & 1/2 & 1/3 & 1/4 & 1/5 \\ 1/2 & 1 & 2/3 & 1/2 & 2/5 \\ 1/3 & 2/3 & 1 & 3/4 & 3/5 \\ 1/4 & 1/2 & 3/4 & 1 & 4/5 \\ 1/5 & 2/5 & 3/5 & 4/5 & 1 \end{bmatrix} = T_{10}^{-1}$$

Elle possède une inverse tridiagonale notée $T_{10}$ symétrique dont on ne connait pas explicitement les valeurs propres. On sait seulement qu'elles sont réelles non nulles et toutes différentes.

$$t_{i,i} = \frac{4i^3}{4\,i^2 - 1} \qquad i \neq N$$

$$t_{N,N} = -\frac{N^2}{2\,N - 1}$$

$$t_{i,i+1} = t_{i+1,i} = \frac{-i\,(i+1)}{2\,i+1} \qquad i \in [1,\ N-1]$$

$$t_{ij} = 0 \qquad |i-j| > 1$$

$$T_{10} = \begin{bmatrix} 4/3 & -2/3 & 0 & 0 & 0 \\ -2/3 & 32/15 & -6/5 & 0 & 0 \\ 0 & -6/5 & 108/35 & -12/7 & 0 \\ 0 & 0 & -12/7 & 256/63 & -20/9 \\ 0 & 0 & 0 & -20/9 & 25/9 \end{bmatrix}$$

4.11 J.H. Wilkinson indique, notamment en [21], une matrice $T_{11}$ tridiagonale dont il est difficile de déterminer les valeurs propres. Elle est d'ordre 21 mais peut être essayé pour un ordre N impair quelconque. Elle est définie par

$$t_{i,i} = \frac{N+1}{2} - i \qquad i \in [1, N]$$

$$t_{i,i+1} = t_{i+1,i} = 1 \qquad i \in [1, N-1]$$

$$t_{i,j} = 0 \qquad |i - j| > 1$$

$$T_{11} \begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 2 \end{bmatrix}$$

4.12 Le même auteur étudie en [22] une matrice tridiagonale $T_{12}$ liée aux polynômes de Lamé, pour laquelle le calcul des valeurs propres est très instable. Il indique notamment des essais réalisés pour l'ordre 13 (avec k = 0,9) où une très faible modification du terme $t_{11,11}$ change profondément certaines valeurs propres.

Cette matrice est définie par :

$$t_{i,i} = 4 (i - 1)^2 (1 + k^2) \qquad i \in [1, N]$$

$$t_{i,i+1} = 2i (2i - 1)$$

$$t_{i+1,i} = - (2N - 2i + 2) (2N + 2i - 5)$$

Remarquons que cette matrice $T_{12}$ n'est pas symétrisable car

$$t_{i,\, i+1} \times t_{i+1,i} < 0 \qquad (k \neq 0)$$

Calcul Analytique des valeurs propres d'une matrice de Wolstenholme et de quelques matrices déduites.

-o-o-o-o-o-o-o-o-o-o-

Soit $D_N(\lambda)$ le déterminant de la matrice $(\lambda I - T_1)$ où $T_1$ est une matrice de Wolstenholme d'ordre N.

$$D_N(\lambda) = \lambda \; D_{N-1}(\lambda) - D_{N-2}(\lambda)$$

Posons $\lambda = u + u^{-1}$

$$D_N(u) = (u + u^{-1}) D_{N-1}(u) - D_{N-2}(u)$$

avec

$D_o(u) = 1$

$D_1(u) = u + u^{-1}$

On en déduit :

$D_N(u) = u^N + u^{N-2} + \text{------} u^{-(N-2)} + u^{-N}$

d'où

$$D_N(u) = \frac{u^{N+1} - u^{-(N+1)}}{u - u^{-1}}$$

Si $\lambda$ est une valeur propre de la matric

$$D_N(u) = 0 \longrightarrow \begin{cases} u^{2(N+1)} = 1 \\ u^2 \neq 1 \end{cases}$$

ce qui est vérifié pour

$$u = \exp\left(\frac{i\,k\,\pi}{N+1}\right) \quad k \in [1, N]$$

$$\boxed{\lambda_k(T_1) = 2 \cos \frac{k\,\pi}{N+1} \quad k \in [1, N]}$$

On en déduit de suite les valeurs propres de la matrice $T_2$ par translation de -2, et en changeant les signes.

$$\lambda_k(T_2) = -\left(-2 + 2\cos\frac{k\,\pi}{N+1}\right) = 4\sin^2\frac{k\,\pi}{2(N+1)} \qquad k \in [1, N]$$

Remarquons que : $\cos\dfrac{k\,\pi}{N+1} = -\cos\dfrac{(N-k+1)\,\pi}{N+1}$

et que la matrice $T_4$ est symétrisable ($BC > 0$). Les valeurs propres de $T_4$ se déduisent alors de celles de $T_1$ par une translation de A et une homothétie de $(BC)^{-1/2}$ sur la matrice symétrique semblable.

Le polynôme caractéristique de la matrice $T_5$ s'écrit

$$P_5(\lambda) = (\lambda - 1)\,D_{N-1} - D_{N-2}$$

où $D_{N-1}$ représente le polynôme caractéristique d'une matrice de Wolstenholme d'ordre N-1. En reprenant le changement de variable $\lambda = u + u^{-1}$

$$P(u) = \frac{(u+1+u^{-1})\,(u^N - u^{-N}) - (u^{N-1} - u^{-N+1})}{u - u^{-1}}$$

$$P(u) = 0 \Longrightarrow u^{2N+1} = 1 \quad , \quad u^2 \neq 1$$

$$u = \exp\left(\frac{i\,2\,k\,\pi}{2N+1}\right) \qquad k \in [1, N]$$

$$\lambda_k = 2\cos\frac{2\,k\,\pi}{2N+1} \qquad k \in [1, N]$$

On détermine de la même manière les valeurs propres de la matrice $T_6$ dont le polynôme caractéristique s'écrit :

$$P_6(\lambda) = (\lambda + 1)^2 D_{N-2} - 2(\lambda + 1) D_{N-3} + D_{N-4}$$

et celles de la matrice $T_7$ dont le polynôme caractéristique de la matrice de Wolstenholme correspondante :

$$\begin{cases} t_{1,1} = 1 \\ t_{i,i} = 0 \qquad i \neq 1, N \\ t_{N,N} = -1 \\ t_{i,i+1} = t_{i+1,i} = 1 \qquad \forall_i \in [1, N-1] \end{cases}$$

s'écrit :

$$P_7(\lambda) = (\lambda^2 - 1) D_{N-2} - 2\lambda D_{N-3} + D_{N-4}$$

Note : La remarque de la page précédente (IV 12) sur l'existence de valeurs propres de signes opposés peut se généraliser à toute matrice tridiagonale ayant des éléments diagonaux égaux :

Théorème

Les valeurs propres de toute matrice tridiagonale telle que $\forall_i : a_{i,i} = a$ ( Tous les termes diagonaux égaux ) vérifient

$\forall_i, \exists_j : \dfrac{\lambda_i + \lambda_j}{2} = a$ (valeurs propres, deux à deux symétriques par rapport à a)

Démonstration

Associons à la matrice tridiagonale A, deux matrices respectivement diagonale D(matrice de translation) et tridiagonale T définies par :

$D_{i,i} = a \ (\Longrightarrow D = a\,I)$

$T_{i,i} = 0$

(ANNEXE)

$$\left.\begin{array}{l} T_{i+1,i} = A_{i+1,\,i} \\ T_{i,\,i+1} = A_{i,\,i+1} \end{array}\right\} \Longrightarrow \quad A = D + T$$

Les matrices T et - T sont semblables, ayant le même polynôme caractéristique, en effet :

$\forall_i$, $T_{i+1,\,i} \times T_{i,\,i+1} = (-T_{i+1,\,i}) \times (-T_{i,\,i+1})$

Elles possèdent alors des valeurs propres opposées et égales, donc symétriques par rapport à zéro.
En effactuant la translation par D, on vérifie bien le théorème annoncé.

Corollaire - Si l'ordre n de la matrice A est impair, celle-ci possède une valeur propre égale à (a).

Remarque : La méthode de Tridiagonalisation de H.R. SCHWARTZ [C.R.Acad. Sc. 1955, p. 15]
transforme une matrice "quelconque" en une matrice semblable et tridiagonale de la forme

$$\begin{bmatrix} 0 & C_1 & & & & \\ 1 & 0 & C_2 & & & \\ & 1 & 0 & \ddots & & \\ & & \ddots & \ddots & C_{N-1} & \\ & & & 1 & A_N & \end{bmatrix}$$

Si la trace de la matrice initiale était nulle (ce que l'on peut toujours réaliser par une translation convenable) le terme $A_N$ est nul. On peut alors appliquer le theorème ci-dessous et en déduire que cette matrice possède des valeurs propres symétriques, ce qui est faux. Cette méthode de tridiagonalisation n'est donc pas applicable en toute généralité.

## CHAPITRE V

## PROGRAMMES ET ESSAIS

Une grande partie des essais ont été réalisés sur la CAB 500 du laboratoire dans des langages divers (ALGOL, PAF ou langage-machine). Des vérifications systématiques lorsqu'elles étaient possibles et de nouveaux essais ont été réalisés en ALGOL sur l'ordinateur IBM 7044 de l'Institut de Mathématiques Appliquées de Grenoble. Le programme ALGOL ci-dessous reproduit un programme essayé sur ce dernier ordinateur. Il utilise la plupart des notions déjà utilisées sur la CAB 500 (sauf notamment celles necessitant une écriture en langage-machine, telle que l'écart relatif) et a conduit à des résultats intéressants pour les matrices indiquées au Chapitre IV. Ces essais ont été réalisés avec des listes de paramètres plus importantes que celles indiquées dans le chapitre IV et notamment pour la matrice T9 d'Eberlein qui s'est révelée être une matrice-test très souple. Il n'est pas possible de reproduire ici ces essais mais seulement de signaler par quelques courbes la qualité des résultats ou les difficultés rencontrées.

PRØCEDURE VPLAG (A, B, N, DRØITE, VP) ;

VALEUR N ; ENTIER N ;

REEL TABLEAU A, B, VP ;

BØØLEEN DRØITE ;

CØMMENTAIRE VPLAG calcule les valeurs propres d'une matrice T tridiagonale par la méthode de Laguerre. Cette matrice T est définie par ses termes diagonaux ($A[I] = T_{i,i}$) et les produits de ses éléments codiagonaux ($B[I] = T_{i,i+1} \times T_{i+1,i}$). N est l'ordre de la matrice. A et B sont alors des tableaux à une dimension de N et N-1 éléments. DRØITE est une variable booleenne qui indique le sens de balayage désiré. VP est un tableau dans lequel seront rangées les valeurs propres calculées dans l'ordre de leur calcul ;

DEBUT

ENTIER I, NVP, ISAUT, NU, MUE ;

CØMMENTAIRE NVP est le numéro de la valeur propre à calculer, ISAUT indique le saut à effectuer, NU est un compteur d'itération ;

REEL BINF, BSUP, X0, X1, X2, X3, MU, CØMP, Q2, Q3, P0, P1, P2, D, CØR, S, EPS ;

CØMMENTAIRE BINF et BSUP sont les bornes calculées par la procédure BORNES, X0, X1, X2, X3, sont les valeurs successives des itérées de Laguerre, MU et CØMP sont des indicateurs de multiplicité, Q2, Q3 sont deux valeurs successives de la distance relative entre deux itérées; P0, P1, P2 sont les valeurs du polynôme et de ses deux dérivées utilisées pour calculer la correction CØR de Laguerre, EPS est un nombre petit utilisé dans les sauts relatifs et absolus ;

REEL TABLEAU C0, C1, C2 [1 : N ];

CØMMENTAIRE Coefficients du polynôme caractéristique et des dérivées premières et secondes ;

BØØLEEN SYM, CRØIT, DR ;

CØMMENTAIRE SYM indique si la matrice est symétrisable ou non, CRØIT est VRAI lorsque l'on s'éloigne d'une solution, FAUX lorsque l'on converge, DR indique si l'on a utilisé une distance relative (DR = VRAI)ou une distance absolue (DR = FAUX) au voisinage de zéro ;

AIGUILLAGE AIG := SAUT 1, SAUT 2, SAUT 3, SAUT 4, SAUT 5 ;

EPS := $10^{-7}$ ;

SYMETRISABLE : SYM := VRAI ;

PØUR I := 1 PAS 1 JUSQUA N-1 FAIRE

SI B [I] <0.0 ALØRS DEBUT SYM := FAUX ;

ALLERA LØCALISATIØN

FIN

SINØN SI B [I] = 0.0 ALØRS ALLERA DECØMPØSITIØN ;

CØMMENTAIRE Une matrice est symétrisable lorsque tous les B [I] sont positifs ;

CØEFFICIENT : CØEF 1(A, B, N, C0) ;

CØEF 2(C0, C1, C2, N) ;

LØCALISATIØN : BØRNES (A, B, N, BINF, BSUP) ;

SI DROITE ALORS X0 := X1 := X2 := BINF

SINON X0 := X1 := X2 := BINF

INITIALISATION : VP [1] := X0 ;
NVP := 1 ;

REPRISE : MU := 1.0 ; NU := 1 ;
CROIT := VRAI; Q2 := 0.0;
ISAUT := 1 ;

ITER

SI SYM ALORS STURM (A, B, N, X2, P0, P1, P2, COMP)

SINON HORNER(C0, C1, C2, N, X2, P0, P1, P2);

LAGN (P0, P1, P2, N, DROITE, COR, SPECIAL) ;

SI COR = 0.0 $\vee$ (DROITE $\equiv$ COR > 0.0) ALORS ALLERA SPECIAL

NOUVELLE ITEREE : X3 := X2 - COR ;

SI SI DROITE ALORS X3 > BSUP $\vee$ X3 < VP [1]

SINON X3 < BINF $\vee$ X3 > VP [1]

ALORS ALLERA SORTIE ;

MULTIPLICITE : D := P1↑ 2 - P0 × P2

SI D ≠ 0.0 ALORS MU := P1↑ 2/D ;

COMMENTAIRE Si le rapport MU est indéterminé, on garde la valeur précédente ;

DISTANCE ENTRE DEUX ITEREES :

Q3 := ABS (X3 - X2) ;

D := ABS (X3) ;

SI D < ABS (X2) ALORS D := ABS (X2) ;

```
SI CRØIT ALØRS
        DEBUT SI Q3 = 0.0 ALØRS
                        DEBUT SI Q2 =0.0 ALØRS
                                DEBUT
                                ISAUT := ISAUT + 1 ;
                                ALLERA AIG |ISAUT|
                                FIN
                        SINØN ALLERA SØLUTIØN
        FIN
        SINØN SI Q3 <Q2 ∧ Q3 <10-1 ALØRS
                        DEBUT CRØIT := FAUX ;
                              ISAUT := 1 ;
                              Q2    := Q3
                        FIN
                        SINØN Q2 := Q3
        FIN
        SINØN SI Q3 = 0.0 ∨ (Q3 > Q2 ∧ DR) ALØRS
                        ALLERA SØLUTIØN
                        SINØN Q2 := Q3 ;

NU := NU + 1 ; X0 := X1 ; X1 := X2 ; X2 := X3 ;
SI NU < 100 ALØRS ALLERA ITER
            SINØN ALLERA FINVPLAG ;
SØLUTIØN :
        Q2 := 0.0 ;
        MUE := ENTIER (MU + 0.5) ;
        VP [NVP] := X2 ;
        NVP := NVP + MUE - 1 ;
```

```
SI  NVP > N  ALØRS  ALLERA  FINVPLAG ;
CØMMENTAIRE  Il y a erreur sur la calcul de la multiplicité;
PØUR I := NVP - MUE + 1  PAS  1  JUSQUA  NVP  FAIRE
                    VP [I ]:= X2 ;
SI NVP = N  ALØRS  ALLERA  FINVPLAG ;
CØMMENTAIRE  Toutes les valeurs propres ont été calculées ;

SI  SI  DRØITE ALØRS  X3  > BSUP ∨ X3 <  VP [1 ]
                SINØN  X3 < BINF  ∨X3 >  VP [1 ]

    ALØRS  ALLERA  FINVPLAG ;

CØMMENTAIRE  On vérifie ici que l'itérée X3 appartient
toujours à l'intervalle de localisation. Dans le cas
contraire, des valeurs propres ont été oubliées ;
NU  := 1 ;
NVP := NVP + 1 ;
SAUT 1 : S := 3 × X3 - 2 × X1 ;
         ALLERA  VERIF ;
SAUT 2 : S := 3 × X3 - 2 × X0 ;
         ALLERA   VERIF ;
SAUT 3 : S := SI  DRØITE  ALØRS  X3× (1 + EPS× SIGNE (X3) )
                          SINØN  X3× (1 - EPS× SIGNE (X3) );
         ALLERA  VERIF ;
SAUT  4 :  S := SI DRØITE  ALØRS  X3 + EPS
                           SINØN  X3 - EPS ;
         ALLERA  VERIF ;
```

```
SAUT  5 :  NU := NU + 1  ;
           ISAUT := 3 ;
           SI NU < 100  ALORS  ALLERA  AIG [ISAUT]
                       SINON  ALLERA  FINVPLAG ;
```

CØMMENTAIRE  On limite le nombre possible de sauts ;

```
VERIF :
SI  SI  DRØITE ≠ S < X3  ALORS ALLERA  SAUTSUIVANT ;
X2 := 3 ;
CRØIT := VRAI ;
Q2  := Q3  :=  0.0  ;
SI  SYM  ALØRS  STURM (A, B, N, X2, P0, P1, P2, CØMP)
         SINØN HØRNER ( C0, C1, C2, N, X2, P0, P1, P2) ;
LAGN (P0, P1, P2, N, DROITE, CØR, SAUT SUIVANT) ;
SI CØR = 0.0 ∨ (DRØITE ≠ CØR > 0.0) ALØRS  ALLERA SAUT SUIVANT
SINØN  SI  ISAUT = 1  ALØRS  ALLERA  REPRISE
                      SINØN  ALLERA  ITER ;
SAUT SUIVANT : ISAUT := ISAUT + 1 ;
               ALLERA  AIG [ISAUT] ;
DECØMPØSITIØN :
```

CØMMENTAIRE La décomposition en deux matrices lorsqu'il existe un B [I] nul n'a pas été inclu dans ce programme ;

```
FINVPLAG  :  FIN ;
```

```
PROCEDURE  COEF (A, B) D'ORDRE : (N) COEFFICIENTS : (D) ;
COMMENTAIRE  COEF calcule les coefficients du polynôme
caractéristique d'une matrice tridiagonale définie
comme pour la procédure VPLAG.  D est un tableau à une
dimension de N éléments - En particulier, D[0], coefficient
du terme de plus haut degré, n'est pas défini, car il
est implicitement pris égal à 1. D[N] est le terme cons-
tant donc P[0] = D[N] ;
        VALEUR  N  ;  ENTIER  N ;
        REEL TABLEAU  A, B, D ;
        DEBUT  REEL TABLEAU  E[0, N - 1] ;
                ENTIER  I, J ;
        E[0] := 1  ;  D[1] := - A[1] ;
        POUR I := 1  PAS 1  JUSQUA N-1  FAIRE
                DEBUT  D[I + 1]:= -A[I+1]×D[I]- B[I]×E[I-1];
                        POUR J := I  PAS -1  JUSQUA 2 FAIRE
                        DEBUT  E[J] :=  D[J] ;
                                D[J] :=  D[J] - A[I+1]×D[J-1]
                                        - B[I]×E[J-2]
                        FIN
                        E[1] := D[1] ;  D[1] := D[1]- A[I+1]
                FIN
        FIN  COEFF  ;
```

PRØCEDURE CØEF 1 (C, N, R) ;

CØMMENTAIRE R est solution d'un polynôme de degré N défini par ses coefficients C[I] avec C[0] = 1, implicitement. CØEF 1 calcule alors les coefficients du polynôme-quotient obtenu en divisant par X - R ;

```
        ENTIER N ;
        REEL TABLEAU C ;
        REEL R ;
DEBUT ENTIER I ;
        C[1] := R + C [1] ;
        PØUR I := 2 PAS 1 JUSQUA N FAIRE
        C[I] := C[I - 1] × R + C[I] ;
```

CØMMENTAIRE L'élément C[N] inutilisé par la suite, est égal au résidu de la division ;

```
        N := N - 1
FIN CØEF 1 ;
```

PROCEDURE CØEF 2 (C, D, E, N) ;

COMMENTAIRE CØEF 2 calcule les coefficients de la dérivée première (appelés D) et de la dérivée seconde (appelés E) d'un polynôme de degré N, défini par ses coefficients C. C, D, E sont des tableaux à une dimension de N éléments. Cette disposition, en imposant implicitement D[0] = E[0] = 0 et explicitement E[1] = 0, permet de considérer les dérivées premières et secondes comme des polynômes de degré N, simplifiant ainsi la procédure de calcul par le schéma de HØRNER ;

```
        VALEUR N ; ENTIER N ;
        REEL TABLEAU C, D, E ;
DEBUT ENTIER I ;
D[1] := N ; D[2] := (N-1) × C[1] ;
E[1] := 0.0 ; E[2] := N × (N - 1) ;
PØUR I := 3 PAS 1 JUSQUA N-1 FAIRE
        DEBUT D[I] := (N - I + 1) × C[I - 1] ;
              E[I] := (N - I + 1) × (N - I + 2) × C[I - 2]
        FIN ;
D[N] := C[N - 1] ; E[N] := 2 × C[N - 2]
FIN CØEF 2 ;
```

```
PROCEDURE BORNES (A, B) D'ORDRE : (N) INFERIEURE : (INF)
                                      SUPERIEURE : (SUP);
```

COMMENTAIRE BORNES calcule un intervalle [INF, SUP] contenant toutes les valeurs propres, par les cercles de Gerschgorin et par la méthode de Laguerre. Les bornes sont définies par l'intersection des intervalles obtenus par ces deux méthodes. On remarquera que cette procédure calcule directement les coefficients C[1] et C[2] du polynôme caractéristique et qu'elle est donc indépendante de la procédure COEF. Le calcul par la méthode de Laguerre n'est pas effectué lorsqu'il conduit, par suite d'erreur, à déterminer la racine carrée d'un nombre négatif. Nous signalerons, dans la conclusion du Chapitre V, la possibilité de ne pas utiliser cette procédure ;

```
        VALEUR N ; ENTIER N ; REEL INF, SUP ;
        REEL TABLEAU A, B ;
DEBUT ENTIER I, J ; REEL X, Y, C1, C2 ;
        REEL TABLEAU MIN, MAX[1 : 2] ;
CERCLES DE GESCHGORIN :
MIN [1] := A[1] - 1.0 ; MAX[1] := A[1] + 1.0 ;
X := ABS ( B[1] ) ;
MIN[2] := A[1] - X ; MAX[2] := A[1] + X ;
POUR I := 2 PAS 1 JUSQUA N-1 FAIRE
POUR J := 1, 2 FAIRE
        DEBUT Y := X + 1. 0
```

```
            SI  MAX[J] < A[I] + Y ALØRS MAX[J] := A[I] + Y;
            SI  MIN[J] > A[I] - Y ALØRS MIN[J] := A[I] - Y ;
            X := ABS (B [I] )
        FIN ;
PØUR J := 1, 2  FAIRE
        DEBUT SI  MAX[J]<A[N] + X ALØRS MAX[J] := A[N] + X;
            SI    MIN[J]>A[N] + X ALØRS MIN[J] := A[N] - X;
            X  :=  1.0
        FIN  ;
INF  :=  SI  MIN[1] <MIN[2] ALØRS MIN[2] SINØN MIN[1] ;
SUP  :=  SI  MAX[1]> MAX[2] ALØRS MAX[2] SINØN MAX[1] ;
METHØDE DE LAGUERRE :
C 1  :=  C2  :=  0.0  ;
PØUR I := 1  PAS 1 JUSQUA  N-1  FAIRE
DEBUT C1 := C1 + A [I] ;
      C2 := C2 + A[I]↑ 2 + 2 × B[I]
FIN
      C1 := - C1 - A[N] ;
      C2 := ( C1↑ 2 - C2 - A[N] ↑2 ) / 2.0  ;
      X := (N - 1) ↑2 × C1 ↑2 - 2 ×N × (N - 1) × C2 ;
      SI X < 0.0  ALØRS  ALLERA  FINBØR ;
      X := RAC 2 (X) ;
      Y := (- C1 + X) / N ;
      X := (- C1 - X) / N ;
      SI  X > INF  ALØRS   INF  :=  X ;
      SI  Y < SUP  ALØRS   SUP  :=  Y ;
      FINBØR : FIN  BØRNES ;
```

```
PRØCEDURE  STURM (A, B, N, X, P1, P2, P3, CØMP) ;
```

CØMMENTAIRE STURM calcule les valeurs P1, P2, P3 du polynôme caractéristique (et de ses dérivées premières et secondes) d'une matrice tridiagonale définie par A et B (voir Commentaire de la procédure VPLAG) d'ordre N, pour une valeur X. On remarquera la manière dont est calculé le compteur CØMP des signes de Sturm et qui pose en fait un problème théorique délicat : Quel est le signe d'une quantité nulle ? ;

```
          VALEUR  N, X  ;  ENTIER N ;
          REEL  X, P1, P2, P3, CØMP ;
          REEL  TABLEAU  A, B ;
DEBUT  ENTIER  I, J  ;  REEL  Y  ;
          REEL  TABLEAU  P [1 : 3,  1 : 3]
P[1, 1 ] := P [2, 2]  := 1.0  ;
P[1, 2 ] := X - A [1]  ;
P[2, 1 ] := P[3, 1] := P[3, 2]  := 0.0  ;
J := SIGNE (P [1, 2] ) ;
CØMP := SI J = 1  ALORS  1.0  SINON
        SI J = 0  ALORS  0.5  SINON  0.0  ;
POUR I := 2  PAS 1  JUSQUA  N  FAIRE
DEBUT   Y := X - A[I] ;
```

```
P[1, 3] := Y × P[1, 2] - B[I - 1] ×P[1,1]  ;
P[2, 3] := Y × P[2, 2] + P[1, 2] - B[I - 1]×P[2, 1];
P[3, 3] := Y × P[3, 2] + 2× P[2,2] - B[I-1]×P[3,1];

POUR  J := 1  PAS  1  JUSQUA  3  FAIRE

DEBUT P[J, 1] :=  P[J, 2] ;
      P[J, 2] :=  P[J, 3] ;

FIN

J := SIGNE (P[1, 1]) ×  SIGNE (P[1, 2])  ;
CØMP := CØMP + (SI J = 1  ALØRS 1.0  SINØN
                SI J = 0  ALØRS 0.5  SINØN 0.0) ;

FIN

P1 := P[1, 2] ; P2  := P[2,2] ;  P3 := P[3,2]

FIN STURM ;
```

```
PRØCEDURE HØRNER (C1, C2, C3, N, X, P1, P2, P3) ;
```

CØMMENTAIRE HØRNER calcule les valeurs P1, P2, P3 des polynômes définis par les coefficients C1, C2, C3, de degré N pour la valeur X. Nous l'utilisons pour calculer la valeur du polynôme et de ses deux dérivées grâce à la disposition des coefficients signalée dans le commentaire de CØEF 2 ;

```
        VALEUR  N, X  ; ENTIER  N  ; REEL X, P1, P2, P3;
        REEL   TABLEAU C1, C2, C3 ;
DEBUT  ENTIER  I ;
        P1 := 1.0  ;
        P2 := P3  := 0.0  ;
        PØUR  I := 1  PAS  1  JUSQUA  N  FAIRE
        DEBUT  P1 := P1 × X + C1[I];
               P2 := P2 × X + C2[I];
               P3 := P3 × X + C3[I];
        FIN
FIN HØRNER ;
```

```
PRØCEDURE  LAGN  (PO, P1, P2, N, DRØITE, R, IMPØSSIBLE) ;
```

COMMENTAIRE LAGN calcule la correction de Laguerre $(X - \mathcal{L}_N(X))$ dans une direction fixée par la variable booleenne DRØITE et affecte à R le résultat. Lorsque le Hessien H est négatif ou lorsque l'on doit diviser par zéro, on impose une correction nulle (R := 0.0) et on sort par l'étiquette IMPØSSIBLE. On interpretera, dans la procédure VPLAG, cette impossibilité de calcul comme une appartenance de la variable X à la zone d'incertitude entourant une solution. Il serait possible aussi de remplacer dans ce cas l'itérée de Laguerre par l'itérée de Newton et de vérifier sa monotonie ;

```
        VALEUR  N  ;  ENTIER  N ;
        REEL  PO, P1, P2, R  ;
        BØØLEEN  DRØITE ; ETIQUETTE  IMPØSSIBLE ;
DEBUT  REEL  G, H ;
        H := (N-1) ×((N-1) ×P1 ↑2 - N ×PO ×P2) ;
        SI H <0  ALØRS  DEBUT  R := 0.0  ;
                                ALLERA  IMPØSSIBLE
                        FIN ;
        G := SI DRØITE = PO >0 ALØRS  P1 - RAC 2 (H)
                                SINØN  P1 + RAC2 (H);
        SI G := 0.0  ALØRS DEBUT R := 0.0 ;
                                ALLERA  IMPØSSIBLE
                        FIN ;
                                R := N ×PO / G ;
FIN  LAGN ;
```

## COMMENTAIRE DE LA FIGURE 1

En abcisse : Les indices des itérées successives calculées par la méthode de Laguerre,

En ordonnée : Les valeurs successives des itérées,

PI signale un point d'inflexion (voir commentaire de la figure 3).

VP ignale une valeur propre calculée.

Cette figure représente le calcul des valeurs propres de la matrice $T_1$ pour N = 6 dont les valeurs propres sont :

$\lambda_1 = -1.801\ 19377$

$\lambda_2 = -1.246\ \ 9796$

$\lambda_3 = -0.445\ 04190$

$\lambda_4 = \ \ 0.445\ 04190$

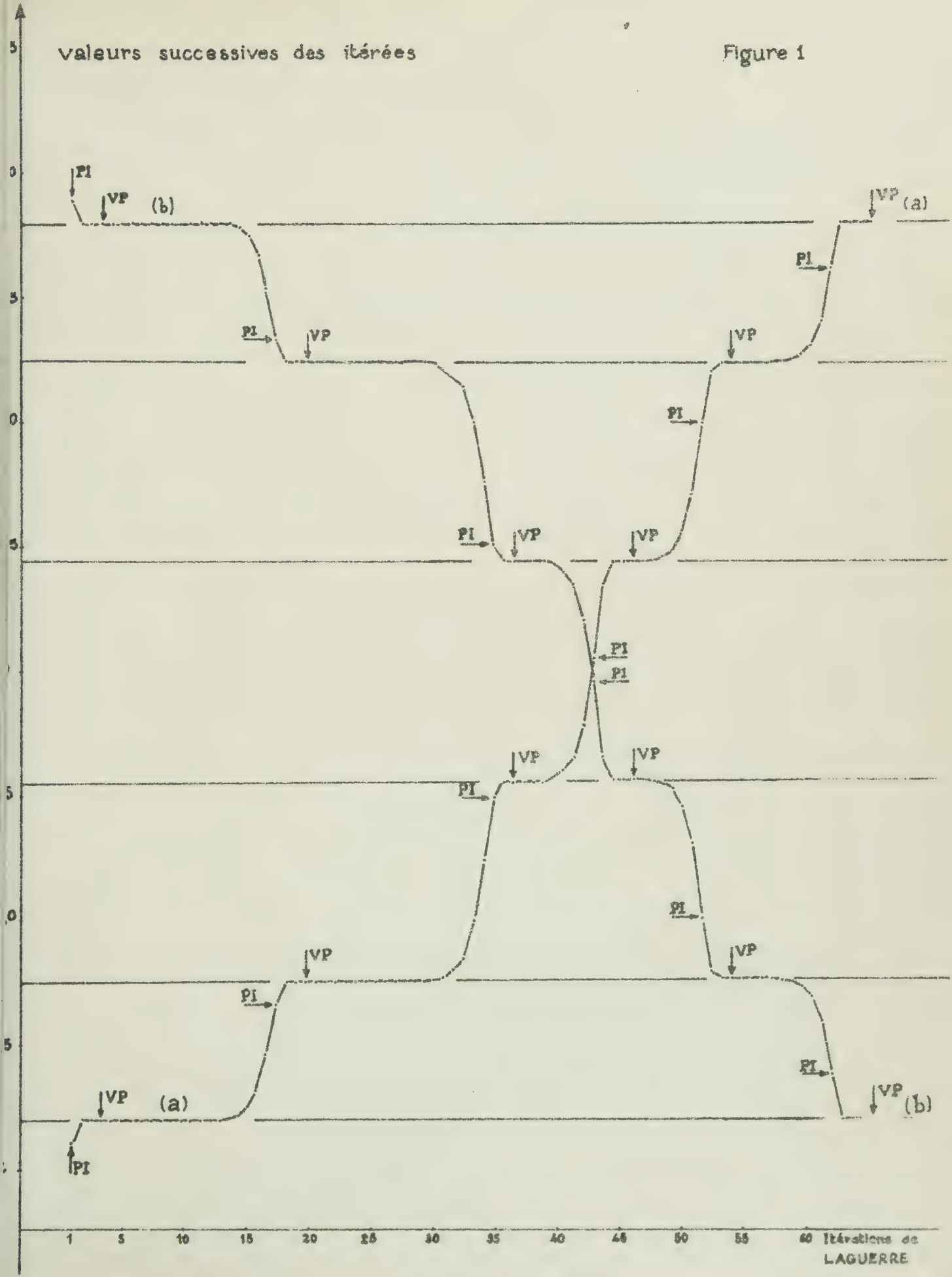
$\lambda_5 = \ \ 0.124\ 69796$

$\lambda_6 = \ \ 0.180\ 19377$

La courbe (a) de bas en haut (respectivement (b) de haut en bas) représente le calcul par un balayage de gauche vers droite (respectivement de droite vers gauche) On remarquera la symétrie de ces deux courbes (a) et (b).

valeurs successives des itérées

Figure 1



## COMMENTAIRE DE LA FIGURE (2)

En abcisse : Les indices des itérées successives calculées par la méthode de Laguerre.

En ordonnée : Les valeurs successives des itérées.

VP signale une valeur propre calculée.

Cette figure représente le calcul des valeurs propres de la matrice $T_9$ pour N = 5 et S = - 20, dont les valeurs propres sont :

$\lambda_1 = 0$

$\lambda_2 = 18$

$\lambda_3 = 34$

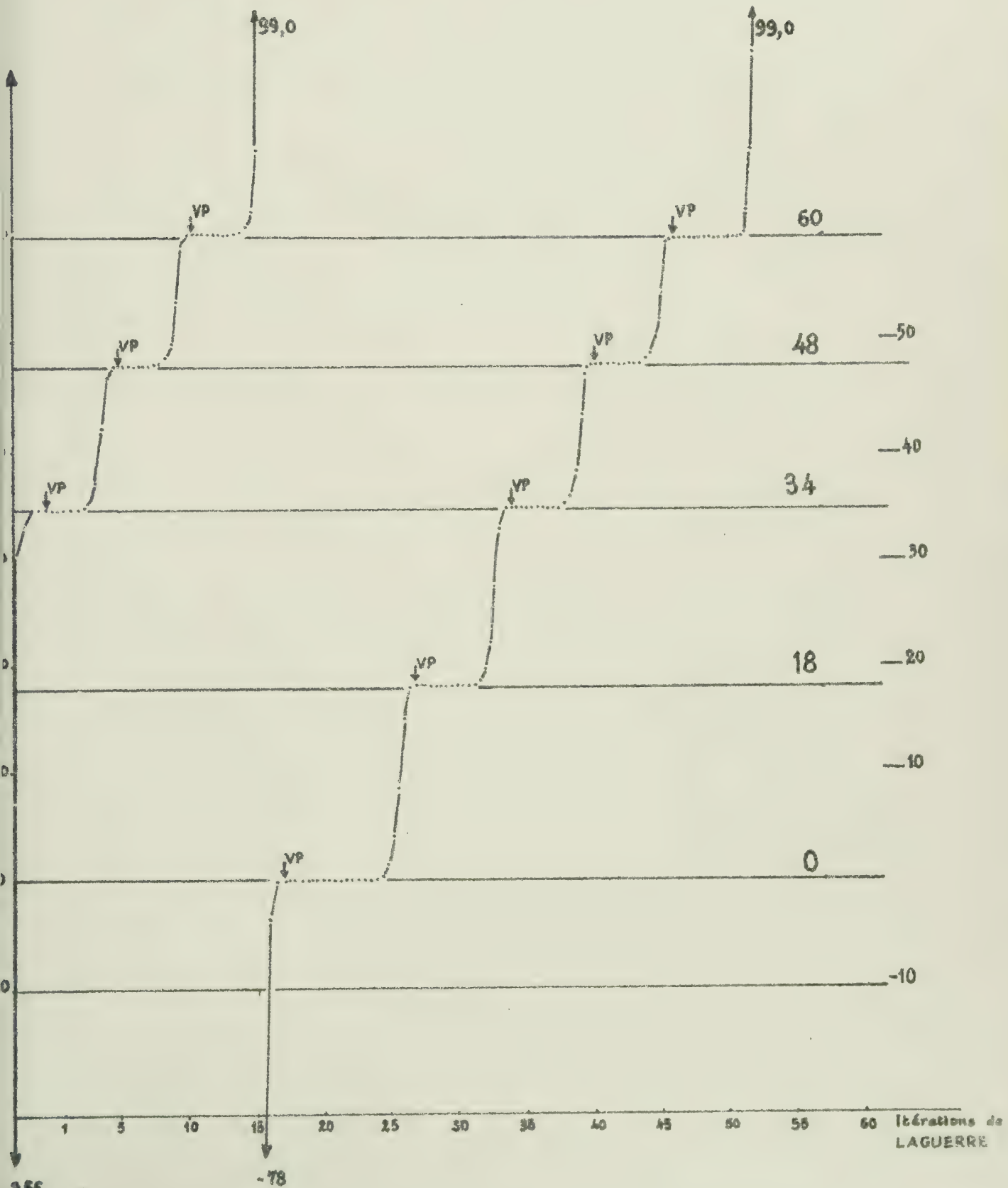
$\lambda_4 = 48$

$\lambda_5 = 60$

On remarquera que le choix d'une valeur initiale trop grande (- 255) a conduit, à cause du dépassement de capacité, à oublier les valeurs propres 0 et 18. Après le bouclage par l'infini (et le "choix" d'une nouvelle valeur initiale - 79) le calcul de toutes les valeurs propres s'effectue normalement ;

valeurs successives des itérées

Figure 2



## COMMENTAIRE DE LA FIGURE (3)

En abscisse : Les indices des itérées successives calculées par la méthode de Laguerre.

En ordonnée : La valeur de l'écart relatif (décimal) entre deux itérées successives.

PI signale un maximum. Ce point est utilisé comme point d'inflexion dans les figures (1) et (2) et sera signalé dans les figures suivantes.

VP signale une valeur propre calculée.

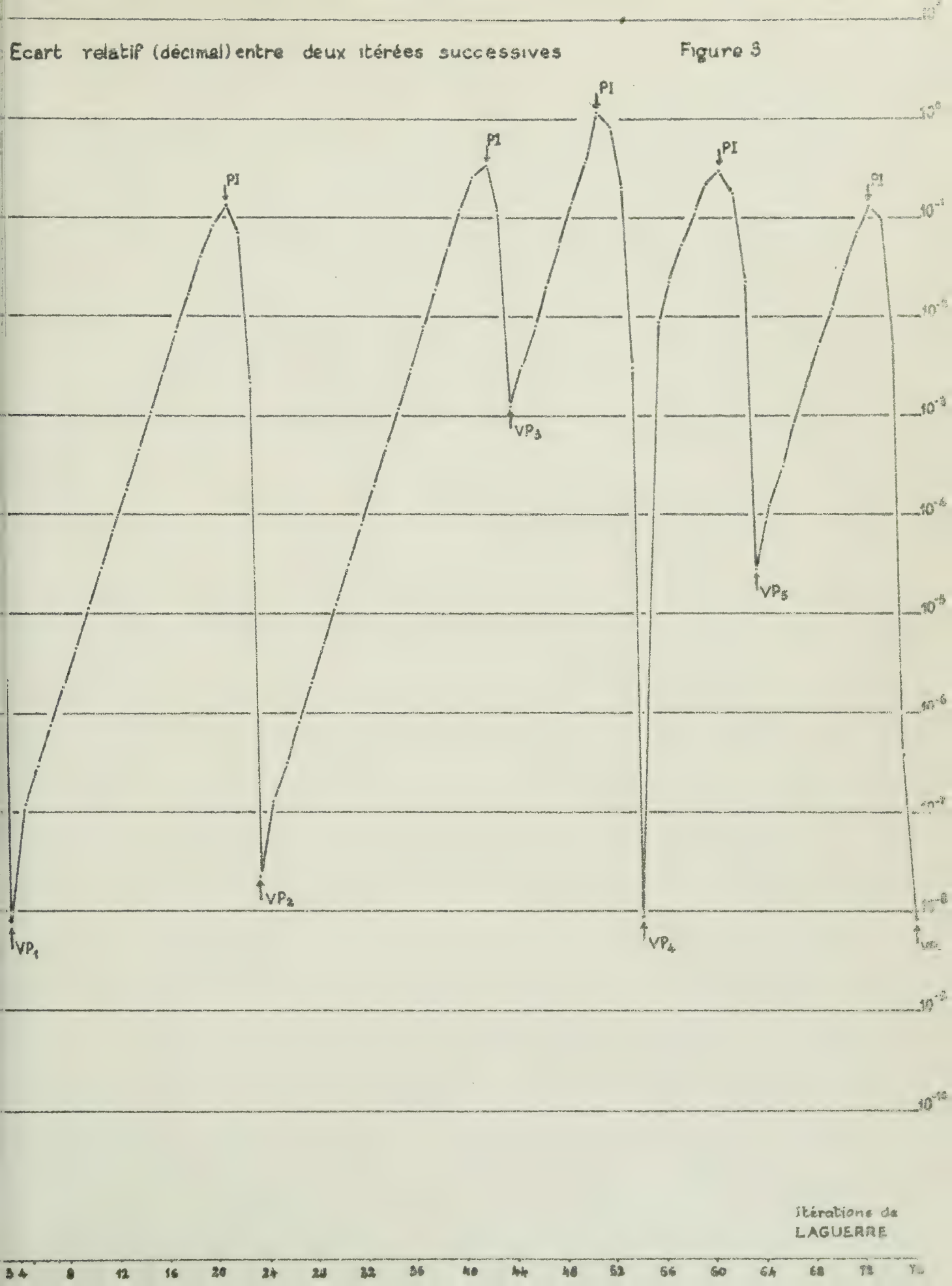
Cette figure représente le calcul des valeurs propres de $T_1$ pour N = 6 déjà utilisée en figure (1).

On remarquera la très rapide convergence vers une valeur propre (nombre de points réduit entre PI et VP) et la lenteur à quitter la solution précédemment calculée (grand nombre de points entre VP et PI)

On constate que les valeurs propres $\lambda_1$, $\lambda_2$, $\lambda_4$, $\lambda_6$ ont été bien calculées (précision $10^{-8}$ environ) et que $\lambda_5$ et surtout $\lambda_3$ sont moins précises ;

Ecart relatif (décimal) entre deux itérées successives

Figure 3



## COMMENTAIRE DE LA FIGURE (4)

En abcisse : Les indices des itérées successives calculées par la méthode de Laguerre.

En ordonnée : Les valeurs de la distance relative entre deux itérées successives.

PI signale un maximum (point d'inflexion)

VP signale une valeur propre calculée.

Cette figure représente le calcul des valeurs propres directement sur le polynôme caractéristique de degré N = 4 dont les solutions sont :

$$\lambda_1 = -1$$

$$\lambda_2 = 0$$

$$\lambda_3 = 0.98$$

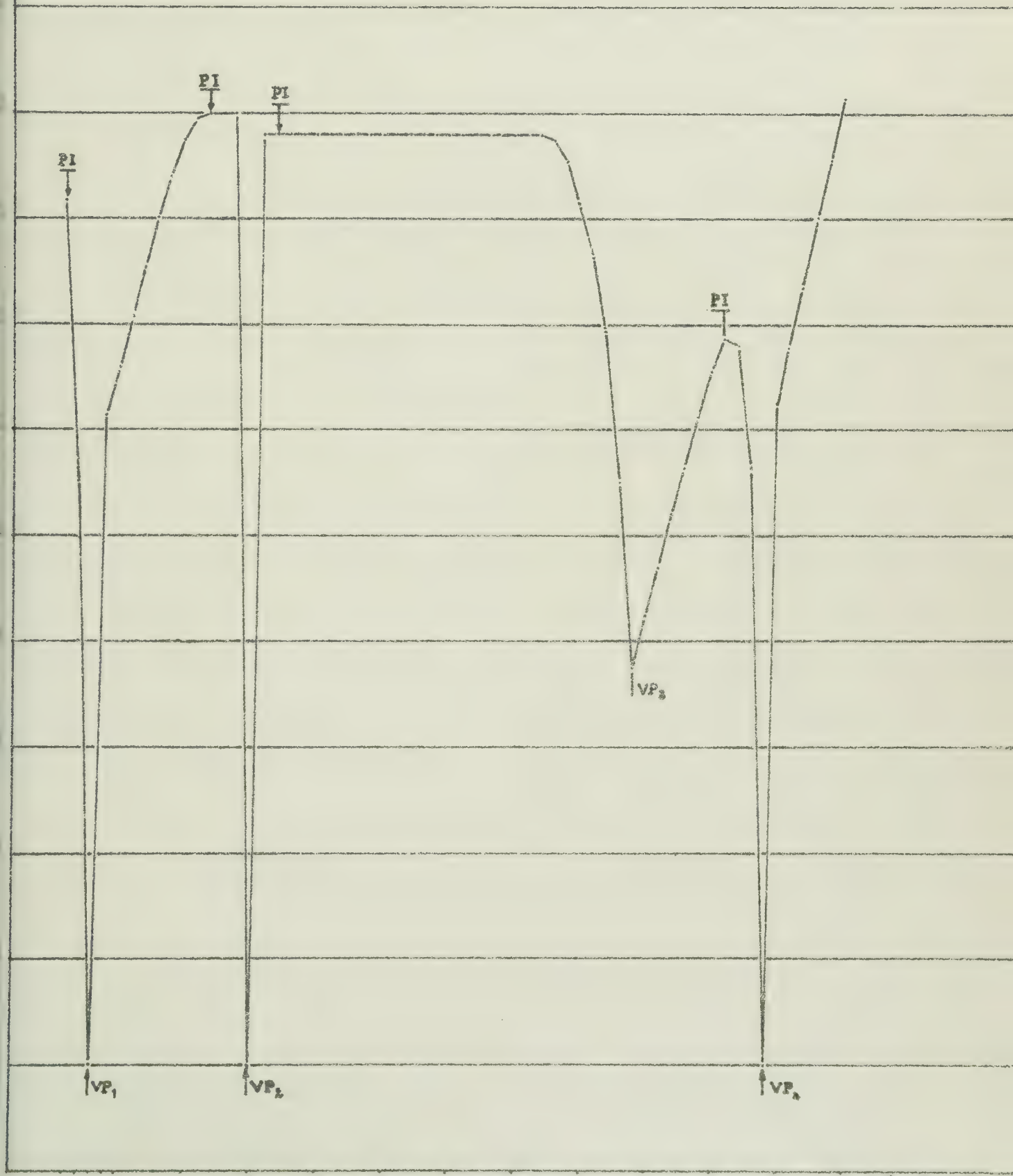
$$\lambda_4 = 1$$

On remarquera le manque de précision (environ $10^{-5}$) de la solution $\lambda_3$ proche de $\lambda_4$, la convergence très spéciale vers la solution $\lambda_2 = 0$ (due à la définition du zéro-machine).

On comparera utilement cette figure avec la figure (5).

Distance relative entre deux itérées successives

Figure 4



## COMMENTAIRE DE LA FIGURE (5)

En abcisse : Les indices des itérées successives calculées par la méthode de Laguerre.

En ordonnée : Les valeurs de l'écart relatif (binaire) entre deux itérées successives.

PI signale un maximum (point d'inflexion)

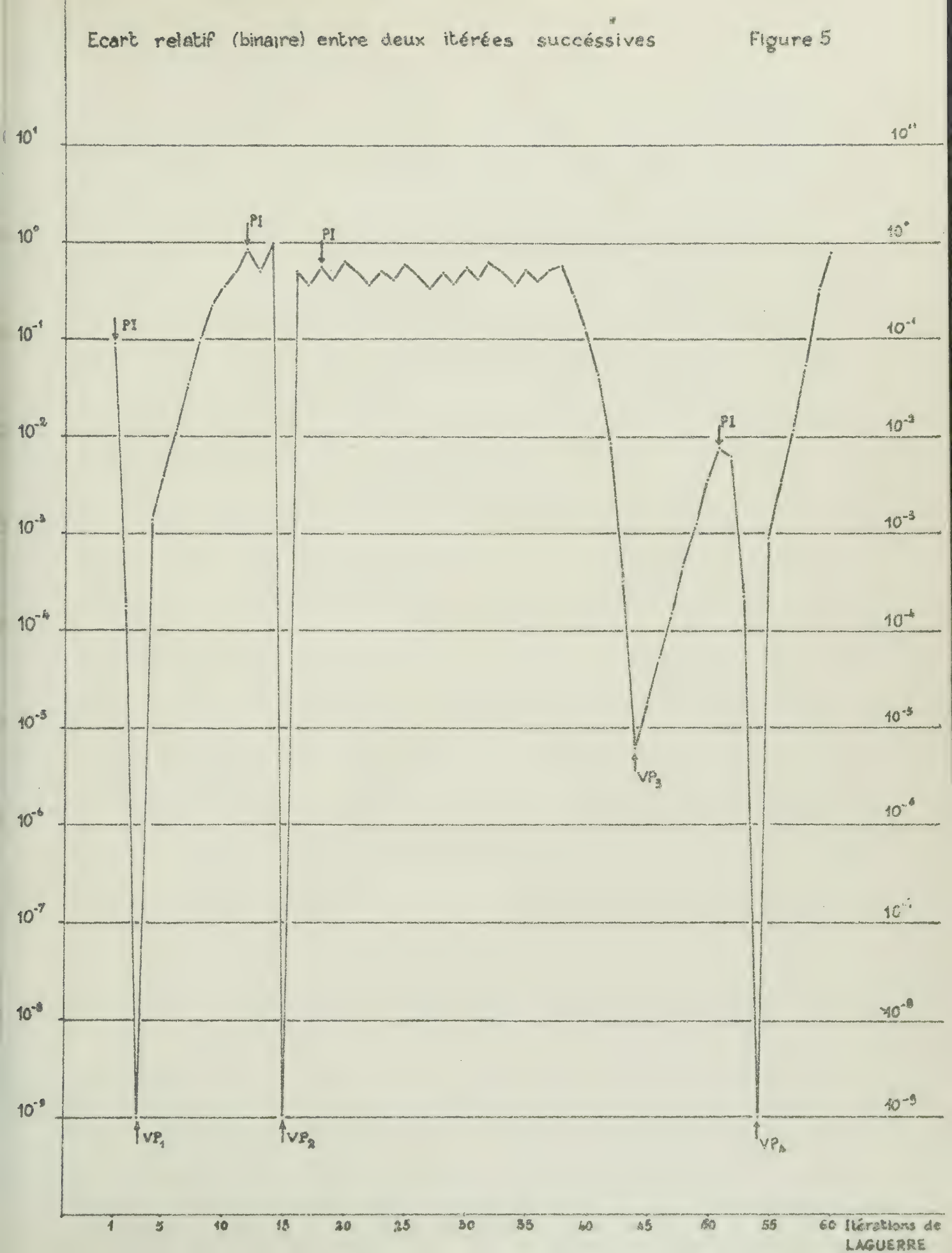
VP signale une valeur propre calculée.

Cette figure représente un calcul analogue à celui de la figure (4) mais en utilisant l'écart relatif (binaire)

On remarquera les oscillations pour les valeurs de l'écart supérieures à $10^{-1}$ et l'identité des figures (4) et (5) pour les valeurs de l'écart inférieures à $10^{-1}$. C'est pour cette raison que nous avons introduit dans la définition de PI, la condition "écart ou distance inférieure à $10^{-1}$".

Ecart relatif (binaire) entre deux itérées successives

Figure 5



## COMMENTAIRE DE LA FIGURE (6)

En abcisse : Les indices des itérées successives calculées par la méthode de Laguerre.

En ordonnée : Les valeurs de l'écart relatif (décimal) entre les valeurs du polynôme caractéristique calculées par les schémas de Sturm et de Horner.

PI signale un point d'inflexion

VP signale une valeur propre calculée.

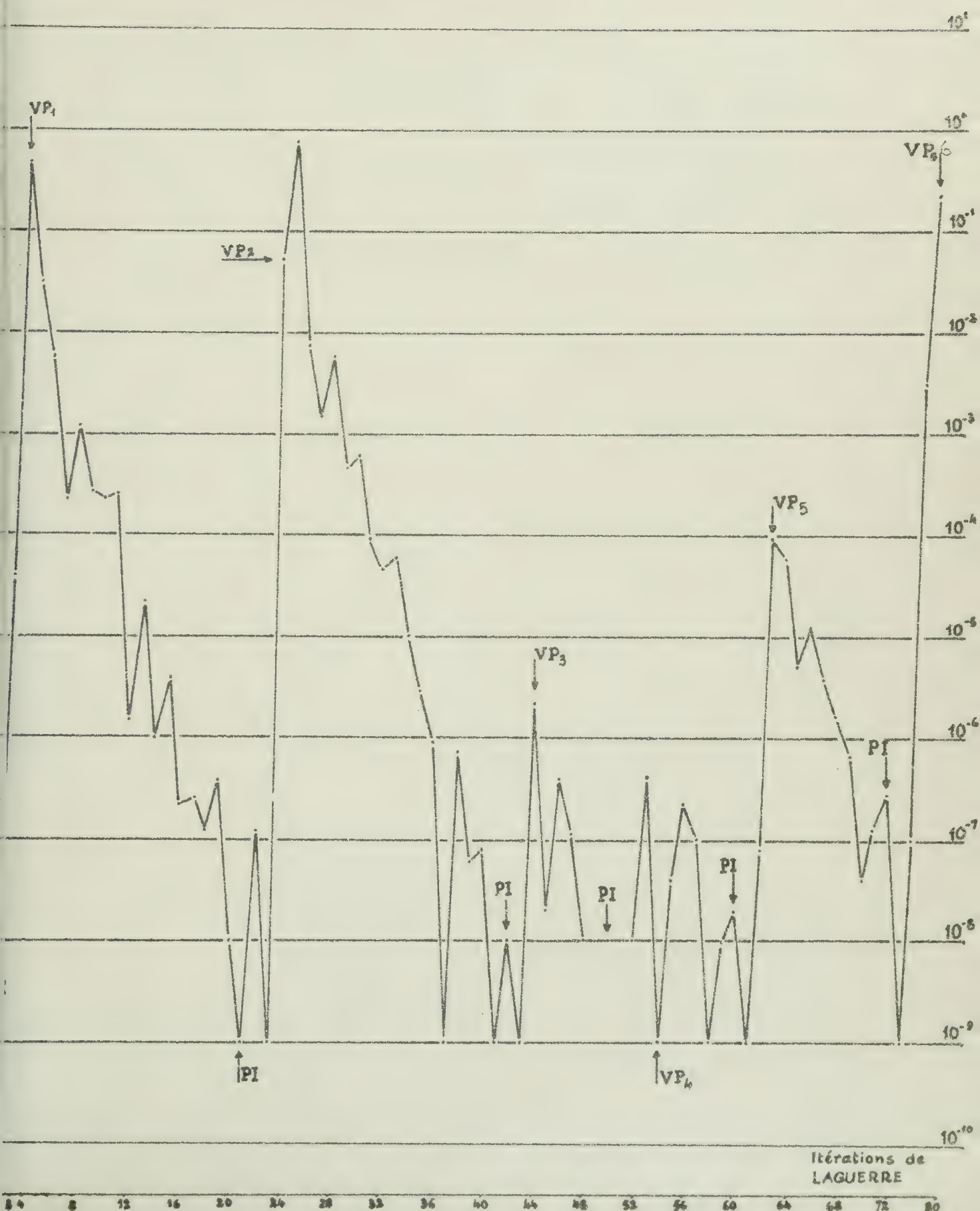
Cette figure représente le calcul des valeurs propres de la matrice $T_1$ pour N = 6 (Voir aussi figure 1 & 3)

On remarquera, en général, une grande différence entre les valeurs calculées par ces deux schémas, au voisinage des valeurs propres et par contre, une bonne précision au voisinage de PI.

Le caractére irrégulier de cette courbe suggère que la loi d'erreur du calcul du polynôme n'est pas une loi simple ;

Ecart relatif (décimal) entre les valeurs du Polynome, calculées par Sturm et Horner

Figure 6



## COMMENTAIRE DES FIGURES (7,1) (7,2) (7,3)

En abcisse : Les indices des itérées successives calculées par la méthode de Laguerre.

En ordonnée : Les valeurs de l'écart relatif (décimal) entre les valeurs du polynôme caractéristique calculées par les schémas de Sturm et de Horner.

PI signale un point d'inflexion.

VP signale une valeur propre calculée.

Comme pour la figure (6), cette courbe représente le calcul des valeurs propres de la matrice T5 pour N = 14 dont les valeurs propres sont :

$\lambda_1 = 1.9532411$ $\quad\quad$ $\lambda_8 = - 0.32356398$

$\lambda_2 = 1.8151508$ $\quad\quad$ $\lambda_9 = - 0.74027628$

$\lambda_3 = 1.5921861$ $\quad\quad$ $\lambda_{10} = - 1.1223741$

$\lambda_4 = 1.2947726$ $\quad\quad$ $\lambda_{11} = - 1.4519909$

$\lambda_5 = 0.93681690$ $\quad\quad$ $\lambda_{12} = - 1.7137143$

$\lambda_6 = 0.53505670$ $\quad\quad$ $\lambda_{13} = - 1.8953063$

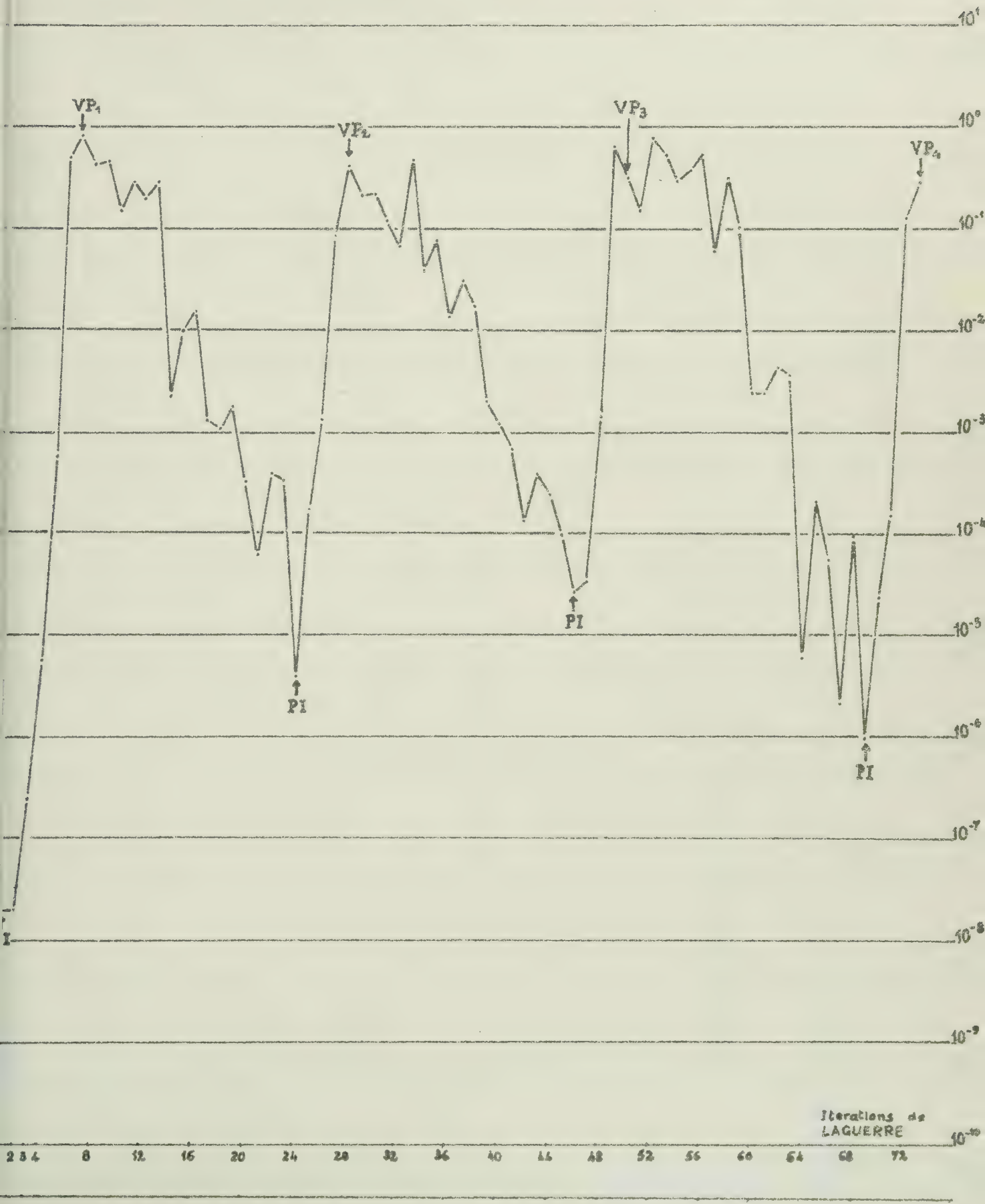
$\lambda_7 = 0.10827785$ $\quad\quad$ $\lambda_{14} = - 1.9882759$

Ces courbes confirment la remarque du commentaire précédent.

On notera cependant le caractère particulier de la valeur propre $\lambda_8$ ;

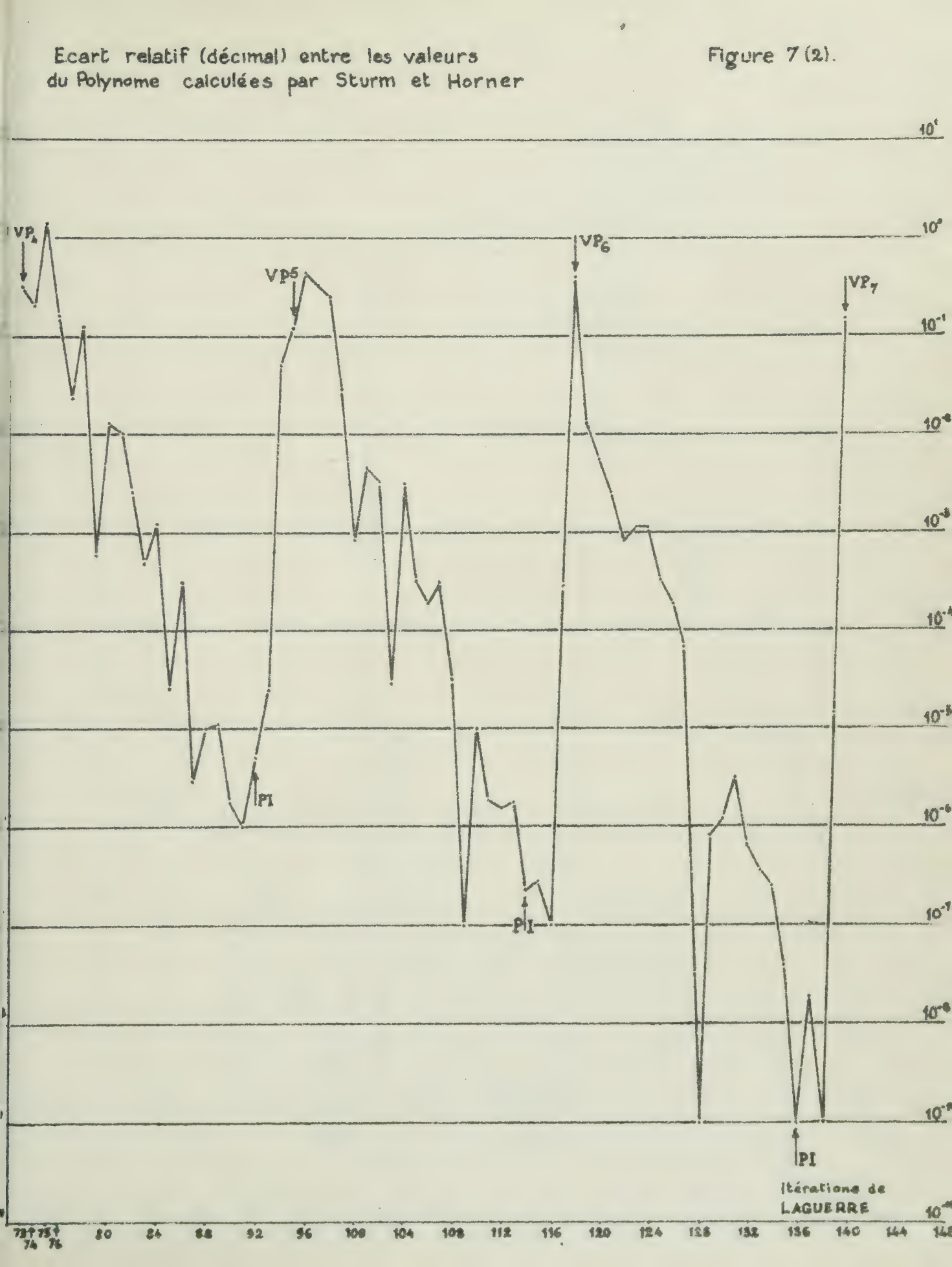
Ecart relatif (décimal) entre les valeurs du Polynome calculées par Sturm et Horner

Figure 7 (1).



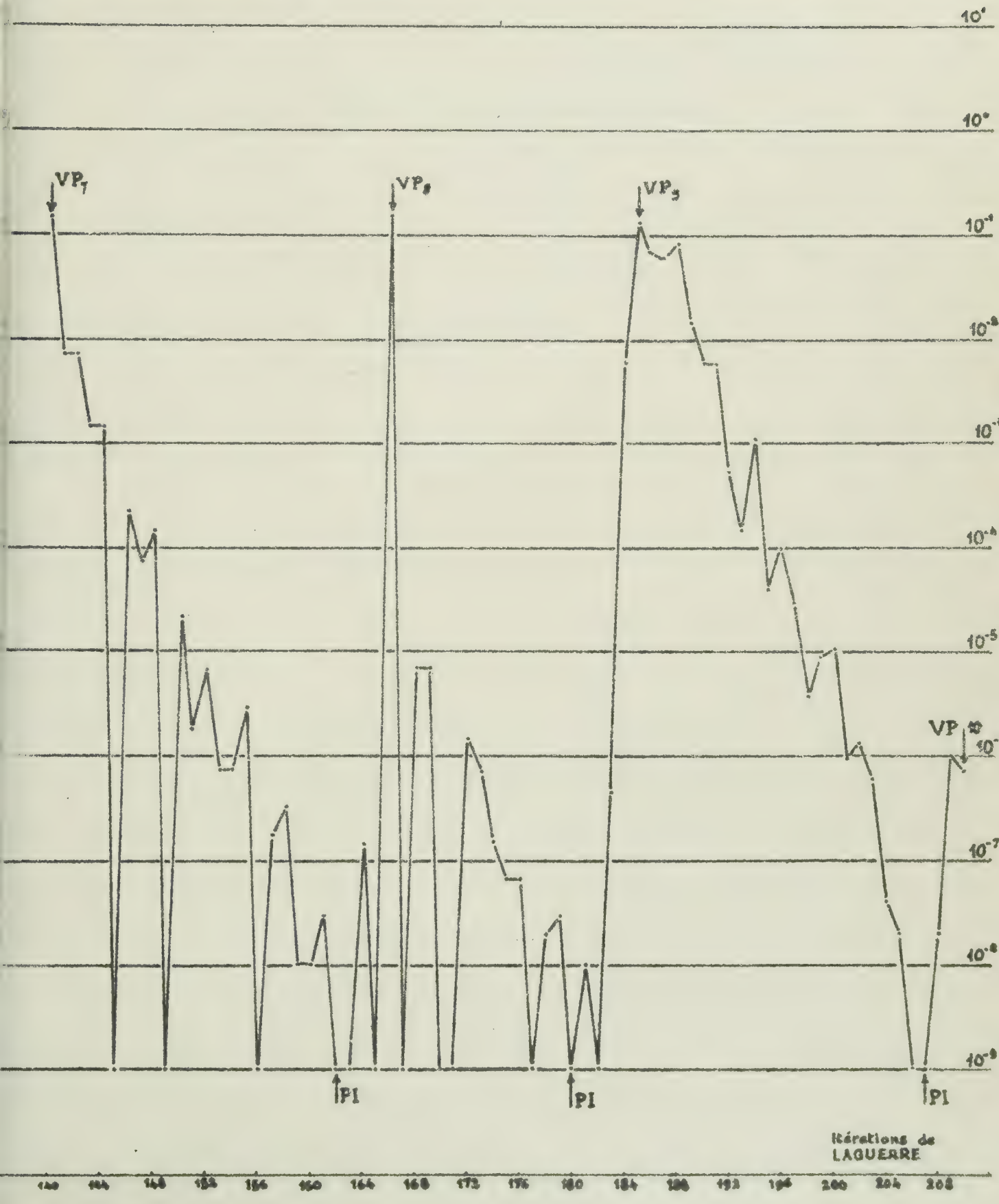
Ecart relatif (décimal) entre les valeurs
du Polynome calculées par Sturm et Horner

Figure 7 (2).



Ecart relatif (décimal) entre les valeurs du Polynome calculées par Sturm et Horner

Figure 7 (3).



## COMMENTAIRE DE LA FIGURE (8)

En abcisse : Les valeurs successives des itérées (axe réel).

En ordonnée : Les valeurs de l'écart relatif (décimal) entre les valeurs du polynôme caractéristique calculées par les schémas de Sturm et de Horner.

PI signale un point d'inflexion.

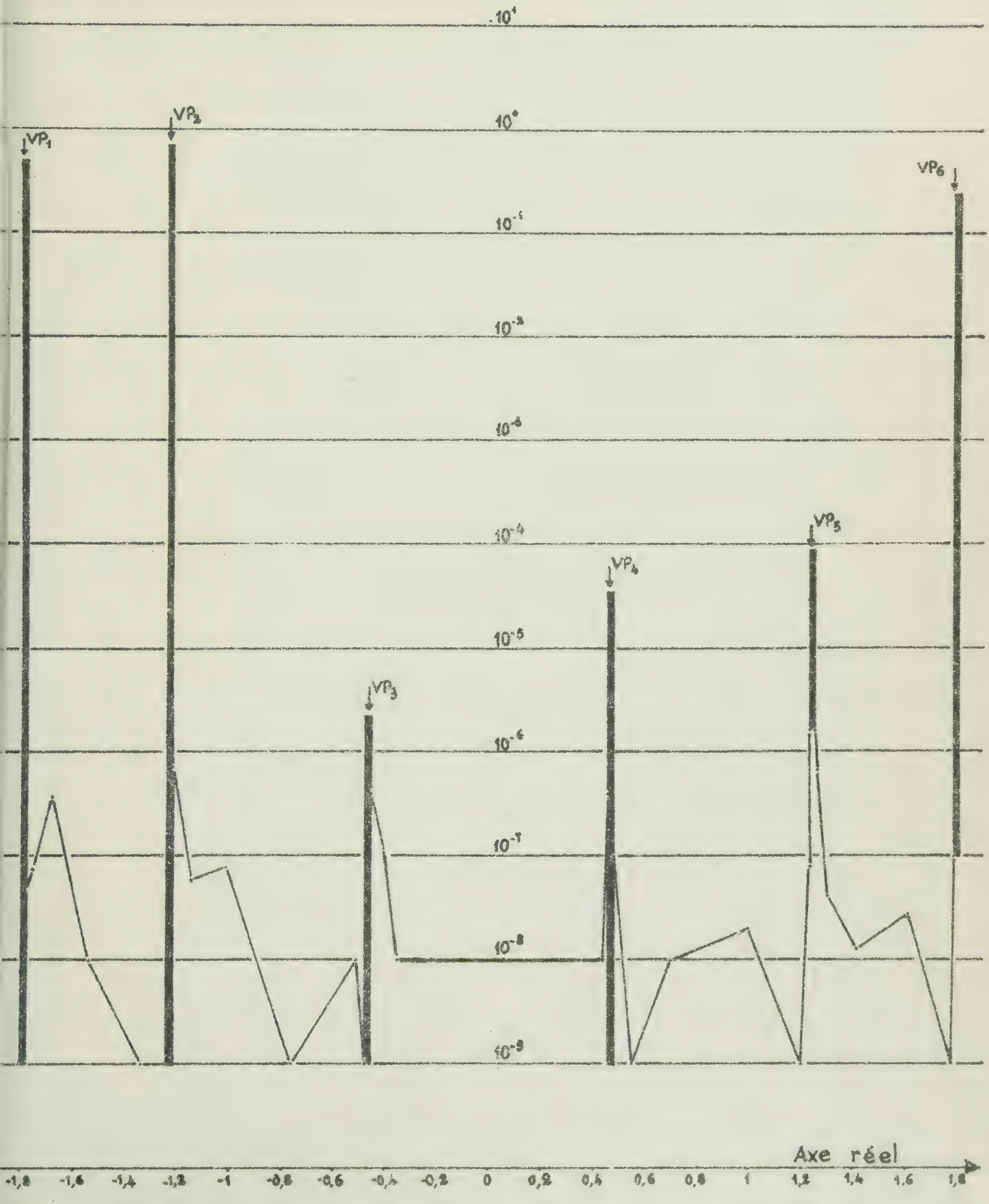
VP signale une valeur propre calculée.

Cette figure représente le calcul des valeurs propres de la matrice $T_1$ pour N = 6 (Voir aussi figures 1, 3, 6).

Les gros traits noirs verticaux matérialisent la convergence au voisinage d'une valeur propre. Leur longueur reste significativecar elle indique le minimum et le maximum de l'écart relatif obtenus et indique donc la précision que l'on peut admettre pour chaque valeur propre ;

Ecart relatif (décimal) entre les valeurs du Polynome calculées par Sturm et Horner

Figure 8



## COMMENTAIRE DES FIGURES (9,1) et (9,2)

En abcisse : Les indices des itérées successives calculées par la méthode de Laguerre.

En ordonnée : Les valeurs successives du rapport $\mu$ .

PI signale un point d'inflexion.

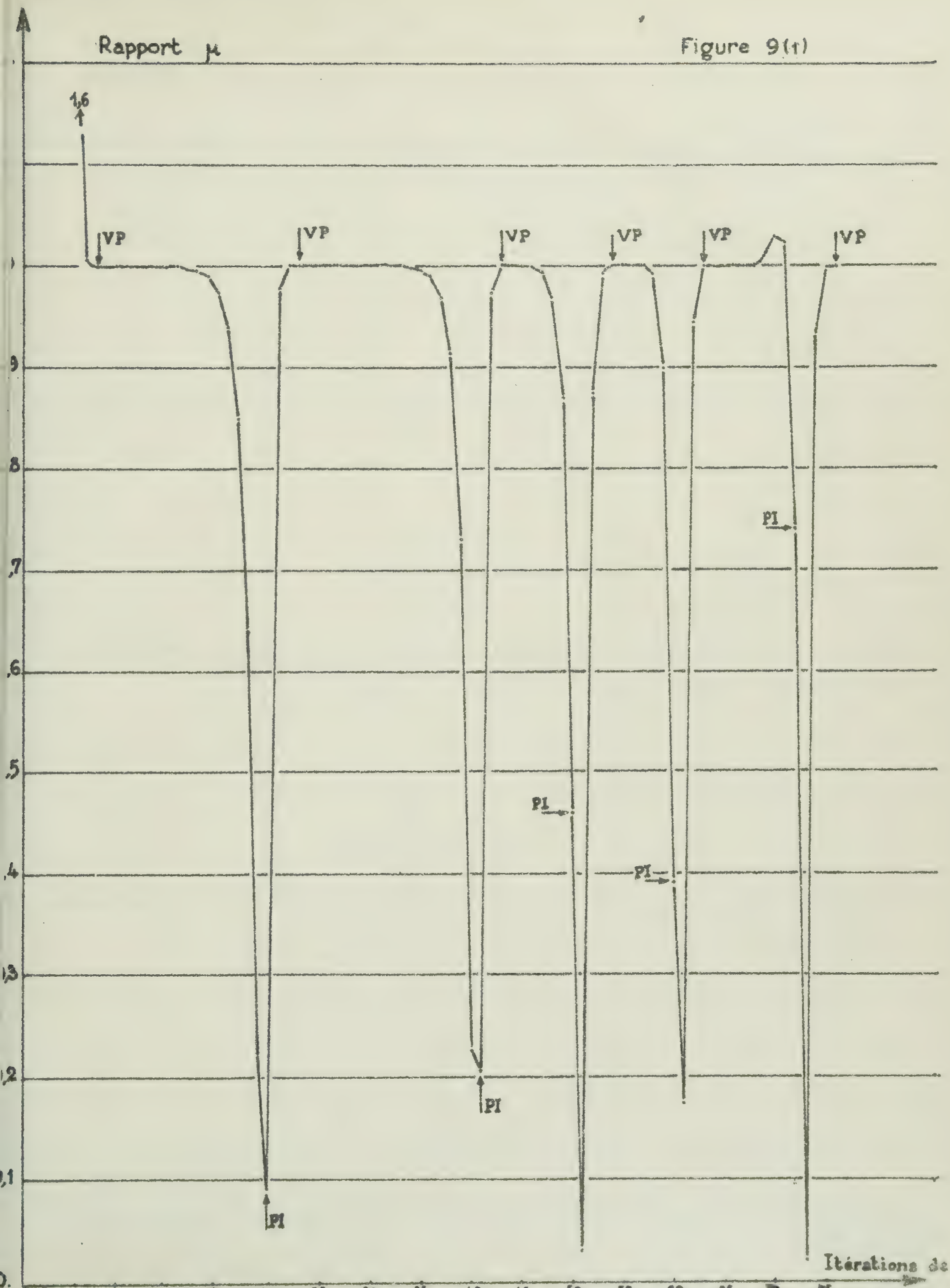
VP signale une valeur propre calculée.

La figure (9,1) représente le calcul des valeurs propres de la matrice $T_1$ pour $N = 6$ (voir aussi figures 1, 3, 6, 8). La figure (9,2) représente le calcul des solutions du polynôme qui a pour solutions -1, -1, 1, 1.

On constate une bonne convergence au voisinage immédiat d'une solution (ou valeur propre) mais nous n'avons pas fait d'essais systématiques.

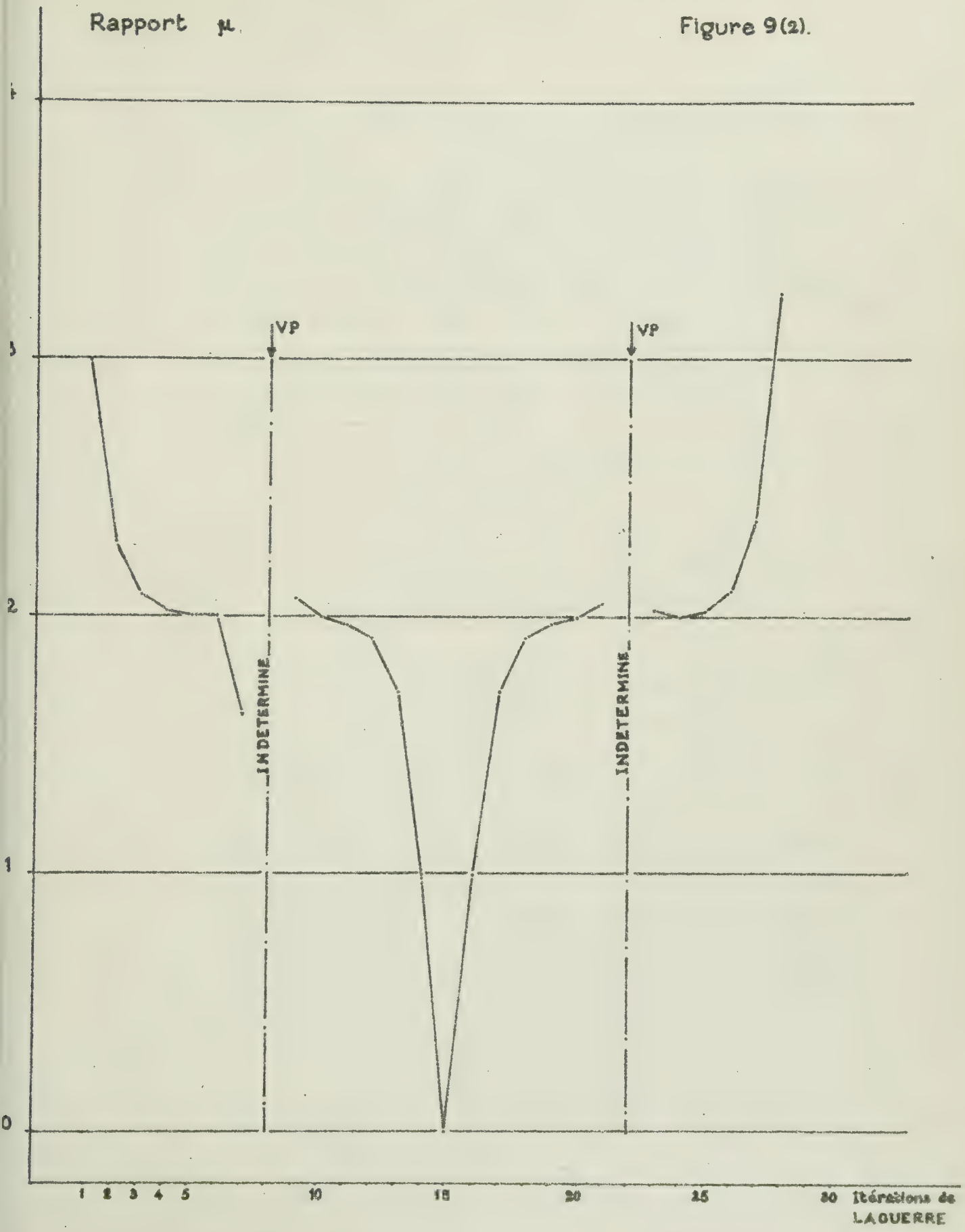
Rapport $\mu$

Figure 9(1)



Rapport $\mu$.

Figure 9(2).



## CONCLUSION

Les essais que nous avons effectués nous ont signalés certaines difficultés auxquelles nous avons essayé de remédier. Notons en particulier le fait que la convergence est toujours très rapide vers une valeur propre mais il est long de s'éloigner de la valeur précédemment calculée (Recherche du point d'inflexion noté PI). Ce point PI possède deux propriétés intéressantes :

- La valeur du polynôme est généralement très précise pour cette valeur.
- Le compteur des signes de Sturm a généralement une valeur exacte donc localise bien les valeurs propres.

Pour les matrices symétriques, il est alors intéressant d'utiliser ces propriétés pour détecter les valeurs propres "oubliées" ou "supplémentaires". D'autres difficultés n'on pu être résolues par cet algorithme. Nous les indiquons en suggérant une possibilité d'y remédier.
Pour éviter les dépassements de capacité et notamment lorsque l'on utilise la procédure VPLAG sur des ordinateurs abandonnant le travail lorsqu'il y a un tel dépassement, on peut effectuer une translation sur la matrice pour ramener toutes les valeurs propres dans l'intervalle $[-1, +1]$

En effet, les valeurs propres étant supposées toutes réelles, les valeurs propres du carré de cette matrice sont alors toutes positives et l'on peut calculer la trace t qui donne alors un intervalle de localisation. Une translation par t/N amène toutes les valeurs propres dans l'intervalle désiré. Dans ces conditions, on peut alors se dispenser de déclarer la procédure BORNES en posant INF:=-1 et SUP:= +1.

La décomposition en sous-matrices lorsque l'un des B[I] est nul est une possibilité théorique car en fait, il faudrait indiquer une valeur minimale au dessous de laquelle serait effectuer la décomposition. Dans ces conditions, cette décomposition n'a pas été introduite dans l'algorithme.

Lorsqu'il est impossible de calculer la formule de Laguerre (Hessien négatif ou division par zéro), il serait souhaitable de déterminer alors l'itérée de Newton si elle est acceptable.

On pourrait d'ailleurs généraliser cette utilisation de la formule de Newton au voisinage immédiat d'une solution.

Les valeurs propres calculées sont rangées dans un tableau à une dimension de N mémoires. Nous avons vu que l'on pouvait estimer la précision des résultats obtenus. Il est donc possible de ranger dans un second vecteur une approximation du diamètre de la zone d'incertitude correspondant à chaque valeur propre. Il serait bon aussi de prévoir dans la tête de procédure, une variable booléenne pour indiquer si une difficulté non résolue a été rencontrée.

Notons enfin qu'il est illusoire de calculer les valeurs propres d'une matrice, en utilisant, implicitement ou explicitement le polynôme caractéristique lorsque celui-ci ne peut être déterminé avec précision. Une réelle amélioration pourrait apportée à cet algorithme en modifiant dans ce sens les procédures Sturm ou Horner car ces procédures conduisent à des valeurs assez éloignées au voisinage d'une valeur propre.

## BIBLIOGRAPHIE


[1] J.L. RIGAL - Séminaires de l'Institut de Statistiques (1965) I.5

[2] N. GASTINEL - Matrices du second degré et normes générales en Analyse Numérique linéaire. Publications scientifiques du Ministère de l'Air. Note technique n° NT 110 (1962) I.22

[3] J.F. MAITRE - Contribution à l'étude de la résolution numérique d'un système linéaire mal conditionné par des stratégies gaussiennes. Thèse 3ème cycle Faculté des Sciences de Besançon (1964) I.22

[4] LINSKI - Calcul de fonctions élémentaires sur calculateur électronique (Livre Russe) I.23

[5] MICROPROGRAMME CAB 500 A, Compagnie des machines BULL CAT N° 1806 (Juillet 1963) I.23

[6] N. METROPOLIS et R.L. ASHENHURST - significant Digit Computer Arithmetic - I.R.E. (1958)

[7] J.H. WILKINSON - Rouding errors in algebraic processes Departement of scientific and industrial research (1963) I.27

[8] Arithmétiques à Chiffres Significatifs - Publication du Laboratoire d'Informatique de Besançon (En préparation) I.27

[9] E.N. LAGUERRE - Oeuvres de Laguerre publiées par l'Académie des Sciences - Paris II.2

[10] J.H. WILKINSON - The algebraic Eigenvalue Problem. Monographs on Numerical Analysis (1965). P 443 II.2

[11] B. PARLETT - Application of Laguerre's method to the matrix eigenvalue problem. Stanford University California Technical Report n° 21 (Mai 1962) II.24

[12] J.M. ORTEGA - On Sturm sequences for tridiagonal matrices. Applied Mathematics and statistics Laboratories. Stanford University California Technical Report n°4 (1960) III.11
et J.A.C.M. Vol 7 n° 3 page 1 (Juillet 1960)

[13] L. NOLIN - Notes Personnelles (1963) III.15

[14] J.H. WILKINSON - The calculation of the Eigenvectors of codiagonal Matrices-Computer Journal Vol 1 - P.90-96 III.15

[15] L. NOLIN - Sur l'élimination d'une valeur propre dans une matrice tridiagonale. C.R. Acad. Sci. Paris (Séance du 11 Février 1963) III.21

[16] D.E. RUTHERFORD - Some continuant determinants arising in physics and chemistry, Proc. Royal Soc. Edinburgh, Vol 62 (1947) p.229-236
Vol 63 (1952) p.236-261 IV.2

[17] L. PARKER - The Operator approach to problems of stability and convergence Scripta Mathematica Studies (1957) Appendice II IV.2

[18] D.E. RUTHERFORD - cf reference 16 IV.5

[19] M. NEWMANN et J. TODD - The evaluation of matrix inversion programm
J. SOC. Indust. Appl. Math. Vol 6
n° 4 (Déc.1958) IV.6

[20] P.J. EBERLEIN - A two parameter test matrix
mathematics of computation.Vol 18
n° 86 (1964) IV.7

[21] J.H. WILKINSON - The evaluations of the zeros of ill-conditioned polynomials - Numerische Mathematik
Part I 150-166 (1959)
Part II 167-180 (1959) IV.10

[22] J.H. WILKINSON - The calculation of the Lamé polynomials. Computer Journal Vol 8
n° 3 (1965) pages 273-277 IV.10

[23] A.S. HOUSEHOLDER - The generation of error in digital computation-O.R.N.L. (1955)

[24] J.H. WILKINSON - Error analysis of floating - point computation-Numer Math 2 p. 319 340 (1960)

[25] R.F. MOORE - interval arithmetic and automatic error analysis in digital computing. Standford University California Tech. Rep. n° 25 (Nov. 1962)

[26] M. LEVY - Quelques remarques sur les erreurs dans la méthode de Lanczos pour la recherche des valeurs propres d'une matrice
Chiffres 4ème année n° 2 (Juin 1961)
p. 87-100

[27] E. GOROG - Essai d'une erreur statistique des erreurs de calcul - Thèse 3ème cycle Grenoble (1961)

[28] H. KANNER - Number base conversion in a significant digit arithmetic-Journal of ACM 12 (1965)

[29] R.L. ASHENHURST et N. METROPOLIS - Radix conversion in an unnormalized arithmetic system Math of Comput. 19 (1965)

[30] W.C. WADEY - Floating - Point arithmetics - Journal of ACM 7 (1960)

[31] E. DURAND - Solutions numériques des équations algébriques - Tomes I et II - Paris

[32] BOLLIET - GASTINEL - LAURENT
ALGOL Manuel Pratique.

JURY : M. CHATELET Président

M. RIGAL } Examinateurs

M. PARIZET }

VU et APPROUVE

Besançon, le

Le Doyen de la Faculté des Sciences

VU, et permis d'imprimer,

Besançon, le

Le Recteur de l'Académie de Besançon